Development of a targeted next-generation sequencing gene panel to investigate recurrent mutations in chronic lymphocytic leukaemia

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy

by

Sozan Qadir Karim Zangana

Dec, 2016

Abstract

Chronic lymphocytic leukaemia (CLL) is a mono-clonal B-cell malignancy characterised by heterogeneous clinical course and response to treatment. Recent studies with whole genome or whole exome sequencing have identified novel recurrent genomic lesions, and associated them with adverse clinical outcome of this disease. Owing to their limited sensitivity, the true incidence of these mutations, subclonal complexity and evolution and their roles in disease progression and treatment resistance are still not quite clear. To gain insight into these issues, I developed a highly sensitive (with an average coverage depth being 2250x and the lowest limit of detection 1%) and robust next generation sequencing (NGS) test using HaloPlex and Ion Torrent PGM to target exons of 15 recurrently mutated genes including *TP53, ATM, SF3B1, PCLO, NOTCH1, LRP1B, SAMHD1, FBXW7, BIRC3, HISTIH1E, XPO1, CHD2, MYD88, POT1* and *ZFPM2* in CLL.

In the initial study using this technique, samples from a cohort of 32 cases with progressive and/or therapy resistant CLL before (n = 10) or after chemotherapy (n = 22) were screened. 87.5% of the patients were found to carry at least one somatic non-synonymous mutation in the first 12 targeted genes (VAF: 2-98%). The most commonly mutated genes were *SF3B1*, *ATM*, *TP53* and *PCLO* identified in 11, 10, 9 and 8 cases, respectively. Mutations in *TP53* and its upstream regulator *ATM* appeared to be dominant over other concurrent gene mutations compared to other genes (P = 0.011). Combining NGS and global SNP array analyses revealed a significant association between genomic aberrations in the ATM-p53 pathway and genomic instability. Moreover, prior exposure to DNA-damaging chemotherapy was associated with the bigger numbers of mutation events and mutated genes, although the increases were borderline significant. These results suggest that ATM-p53 pathway defects contribute to the acquisition of additional genomic aberrations and that treatment with DNA-damaging chemotherapy may play a role in the induction or selection of mutations.

In the subsequent longitudinal study of 33 additional samples of 23 mutated cases from the same cohort, I showed the existence of different mutational processes operative in CLL including mutations related to AID, ageing and other factors. I confirmed the significant

contribution of AID-related mutations to CLL clonal evolution, implying on-going activity of this enzyme in off-target genes. In addition, I demonstrated the predominance of a linear path of clonal evolution in this cohort. Thus, 89.3% of the mutated clones/subclones identified at advanced disease stages were detectable at time of diagnosis or prior to treatment. Hence, the early detection of these mutations may potentially serve as predictive biomarkers to inform on therapeutic decisions. I also documented convergent clonal evolution with priori-selection of clones carrying deleterious mutations in 2 target genes including *ATM* and *TP53*. This observation suggests that not all the mutations in the same gene play an equal role in disease progression and/or treatment resistance. Analysis of mutation doubling time revealed that driver mutations, including those in *TP53*, *BIRC3*, *NOTCH1* and *SF3B1,* were significantly correlated with faster evolution as indicated by the shorter doubling time of variant allele frequency. Importantly, I found that increased subclonal sizes were strongly associated with shorter treatment-free survival in patients with driver mutations.

Taken together, the results from this thesis have provided strong evidence emphasising the importance and usefulness of applying the ultra-sensitive NGS test in the early detection and subsequent monitoring of these recurrent somatic mutations in CLL. A translational study based on findings in this thesis is now ongoing, with an aim to convert this test into a regional clinical service.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my primary supervisor Dr. Ke Lin for his endless support, motivation, immense knowledge and his valuable advice that he gave me throughout the course of my study especially in writing of this thesis. Without his precious support, it would not be possible to conduct this research.

I extend sincere thanks to my co-supervisor Prof. Andrew Pettitt for providing me an opportunity to join his research group and all his insightful comments for constructing this work. I am very grateful to my third supervisor Dr. Gillian Johnson for teaching me the basic molecular techniques and for her kind guidance during this work.

I liked to thank Ms. Una Maye and Ms. Emma Price from the Merseyside and Cheshire Regional Genetic Laboratory for their collaborative effort while conducting this study, Ms. Katie Bullock from GCLP Lab for her technical support and Dr. Gareth Weedall from the Centre of Genomic Research, University of Liverpool for his support in NGS data analysis. Besides, I thank all the members of the Department of Haematology and my colleagues Mr. Faris Tayeb, Dr. Umair Khan, Dr. Omar Alishlash, Dr. Marianne Johnson, Mr. Moses Lugos, Mr. Jehad Alhmoud, Mr. Venkateswarlu Perikala, Mr. Ishaque Mohammad, Mr. Ahmed Alshatti, Ms. Sofia Karatasaki, Ms. Faten Yasen, Dr. Ola Alsanabra and Dr. Fatima Talab for their cooperation and encouragements. I also like to thank the Ministry of Higher Education of Iraqi Kurdistan Region for funding this project and Dr. Rekawt Rashid, Dr. Miran Abas, Prof. Anwar Sheikha, Prof. Michael Hughson, Dr. Hisham Al-Rawi, Dr. Runak Majeed, Dr. Heshu Muhammad and Dr. Dosti Najat from the Ministry of Health for their kind support. Special thanks should go to my parents Mrs. Naela Ahmad and Mr. Qadir Karim, all of my sisters and brothers who inspire me as a hardworking person. I am also grateful to my parents in-law Mrs. Maliha Salim and Mr. Rashid Karim and the rest of their family for their continuous support.

Last but not the least; my great appreciation must go to my husband Dr. Faruq Rashid and my sons Mr. Mohammad and Mr. Ahmed and my daughter Miss. Saya for their limitless patience and their encouragement since the start of this study.

Contents

Declaration

I hereby, declare that all of the data presented in this thesis is a result of my own work except the FASAY assay which had been performed by Dr. Gillian Johnson, the copy number study performed by our collaborator from the Merseyside and Cheshire Regional Genetic Laboratory and the Ion Torrent sequencing which was performed by Ms. Katie Bullock from the GCLP laboratory, University of Liverpool.

Publications arising from this work

S. Karim (Blood 2014) 124 (21). Development of an Ultra-deep Sequencing Method on Ion Torrent PGM to Detect Mutations in a Panel of Genes Relevant to Disease Progression and Drug Resistance in Chronic Lymphocytic Leukaemia.

Presentations arising from this work

- Sozan Karim, Gillian Johnson, Andrew Pettitt and Ke Lin. A longitudinal study of NGS in CLL using Ion Torrent PGM shows evidence of chemotherapy-induced mutagenesis and a snowballing effect due to loss of p53/ATM-mediated DNA repair (Poster presentation) 2015 EHA Congress, Vienna, Austria.

- Sozan Karim, Gillian Johnson, Andrew Pettitt and Ke Lin. Next generation sequencing in CLL using Ion Torrent PGM shows evidence of chemotherapy-induced mutagenesis and a snowballing effect due to loss of p53/ATM-mediated DNA repair (Poster presentation) 2015 BSH Conference, Edinburgh, UK.

- Sozan Karim, Gillian Johnson, Andrew Pettitt and Ke Lin. A longitudinal study of NGS in CLL using Ion Torrent PGM shows evidence of chemotherapy-induced mutagenesis and a snowballing effect due to loss of p53/ATM-mediated DNA repair (Oral presentation) 2015 Agilent Genomics User Meeting, London, UK.

- Sozan Karim, Gillian Johnson, Andrew Pettitt and Ke Lin. Development of an Ultra-deep NGS Method on Ion Torrent PGM to Detect Mutations in a Panel of Genes Relevant to Disease Progression and Drug Resistance in Chronic Lymphocytic Leukaemia (Poster presentation) 2014 Life Technologies User Meeting, Manchester, UK.

List of figures

List of tables

List of abbreviations

A	Adenine
Ab	Antibody
ABT-199	Venetoclax
ACAA1	Acetyl-CoA Acyl-Transferase 1
ADAM17	A Disintegrin and Metaliprotinase Domain 17
ADCC	Antibody Dependent Cell Cytotoxicity
AID	Activation Induced Cytidine De-Aminase
ANK	Ankyrin
APOBEC3	Apo Lipoprotein B mRNA Editing Enzyme Catalytic Polypeptide-Like 3G
AS	Allele Specific
ASXL1	Additional Sex Combs Like 1
ATM	Ataxia Telengectesia Mutated
Av.	Average
BAM	Binary Alignment Map
BCL2	B-Cell Lymphoma2
BCL3	B-Cell Lymphoma 3
BCR	B- Cell Receptor
BED	Browser Extensible Data
BIRC3	Baculoviral IAP Repeat-Containing 3
BME	β-Mercaptoethanol
β2-M	β2-Microglobulin
bp	Base pair
BRAF	V-Raf Murine Sarcoma Viral Oncogene Homolog B1
BSA	Bovine Serum Albumin
BTK	Bruton Agammaglobulinemia Tyrosine Kinase
BWA	Burrow Wheeler Aligner
C	Cytosine
^{o}C	Celsius

Ca^{++}	Calcium
c-AID	Canonical-AID
CAL101	Idelalisib
CNA	Copy Number Aberrations
CARD	Caspase Recruitment Domain
CAV1	Caveolin 1
CDC	Complement Dependent Cytotoxicity
c.DNA	Complementary DNA
CGH	Comparative Genomic Hybridisation
CHD2	Chromodomain Helicase DNA Binding Protein 2
chr.	Chromosome
Clb	Chlorambucil
CLL	Chronic Lymphocytic Leukaemia
c-MYC	Avian Myelocytomatosis Virus Oncogene Cellular Homolog
CNN-LOH	Copy Number Neutral Loss of Heterozygosity
COSMIC-65	Catalogue of Somatic Mutations in Cancer
Cov.	Coverage
CpG	Cytosine Guanine (Linear dinucleotide)
CSR	Class Switch Recombination
CXCR4	CXC Receptor 4
dbSNP	Database of Single Nucleotide Polymorphism
DD	Death Domain
DDX3X	Dead Box RNA Helicase
del	Deletion
DLEU 1	Deleted in Leukaemia 1
DLEU2	Deleted in Leukaemia 2
DGGE	Denaturing Gradient Gel Electrophoresis
DHLPC	Denaturing High Performance Liquid Chromatography
DLBCL	Diffuse Large B cell Lymphoma

DMSO	Dimethyl-Sulph-Oxide
DNA	Deoxy-Ribo-Nucelic Acid
dNTP	Deoxy-Ribo-Nucleotide Triphosphates
ds	Double Strand
Dx	Diagnosis
EB	Elution Buffer
EDTA	Ethylene-Di-Amine-Tetra-Acetic Acid
EFEMP1	Human EGF-Containing Fibulin-Like Extracellular Matrix Protein 1
ERIC	European Research Initiative on CLL
EtBr	Ethidium Bromide
EZH2	Enhancer of Zeste Homolog 2
F	Female
FASAY	Functional Analysis of Separated Allele in Yeast
FASTQ	File of Sequence Information and Quality Score
FAT1	Human Proto-Cadherin Fat1
FBXW7	F Box and WD Repeat Domain
FFPE	Formalin Fixed Paraffin Embedded
FISH	Fluorescence Insitu Hybridisation
Flu	Fludarabine
FOG	FOG Family Member
fs*	Truncating Frame Shift Mutations
FUBP1	Far Upstream Element Binding protein 1
Fw	Forward
G	Guanine
g.DNA	Genomic DNA
GA-101	Obinutuzumab
GAPDH	Glyceraldehyde-3-Phosphate Dehydrogenase
GATA	GATA Family Transcription Protein
Gbp	Giga-Base pair
H chain	Heavy Chain
H^+	Hydrogen Ion

HD	Histidine-Aspartic Domain
HGF	Hepatocyte Growth Factor
HIST1H1E	Histone Cluster 1
hr.	Hours
HTRA	HTR Protease A
IARC	International Agency for Research on Cancer
IBM SPSS	International Business Machines Statistical Package for Social Sciences
ID	Identification Code
Ig	Immunoglobin
IGHV	Immunoglobin Heavy Chain Variable Genes
IGV	Integrative Genomic Viewer
IKZF3	Ikaros Family Zinc Finger Protein 3
IL1B	Interleukin 1 Beta
indels	Small Insertions or Deletions
ING3	Inhibitor of Growth Family Member 3
IRF4	Interferon Regulatory Factor 4
ISP	Ion Sphere Particles
ITGA6	Integrin Subunit Alpha 6
IWCLL	International Workshop on Chronic Lymphocytic Leukaemia
Kbp	Kilo Base Pair
KLHL6	Kelch Like Family Member 6
KRAS	Kirsten Rat Sarcoma Viral Oncogene Homolog
L chain	Light Chain
LDH	Lactate Dehydrogenase
LDT	Lymphocyte Doubling Time
LN	Lymph Node
LOD	Limit of Detection
LOH	Loss of Heterozygosity
LRP1B	LDL Receptor Related Protein1 B
M	Male
MAP2K1	Mitogen-Activated Protein Kinase 1

MAPK	Mitogen Activated Protein Kinase
Mbp	Mega Base Pair
MDM2	Murine Double Minute 2
MDR	Minimal Deleted Region
MED12	Mediator Complex Subunit 12
µg	Micro Grams
µl	Micro Litre
mg	Milli Gram
MGA	MAX Dimerisation Protein
M-IGHV	Mutated IGHV
min.	Minutes
MIP	Molecular Inversion Probe
miR-15a	Micro RNA 15a
miR-16-1	Micro RNA 16-1
ml	Millilitre
mRNA	Messenger RNA
MTUS1	Microtubule Associated Tumour Suppressor 1
Mu	Mutated
MW	Molecular Weight
MYCN	V-Myc-Avian Myelocytomatosis Viral Oncogene Neuroblastoma
MYD88	Myeloid Differentiating Antigen 88
MΩ	Milli Q
NA	Not Available
NaOH	Sodium Hydroxide
NCBI	National Centre for Biotechnology Information
NDRG1	N-Myc Downstream Regulated 1
NES	Nuclear Export Signals
NFκB1	Nuclear Factor Kappa B Subunit 1
NFκBIE	NFκB Inhibitor Epsilon
NGS	Next Generation Sequencing
nM	Nano Molar

NOTCH1	Human NOTCH1 Gene
NR	No Response
NRAS	Neuroblastoma RAS Viral Oncogene Homolog
NT5E	5- Nucleotidase Ecto
OMIM	On line Mendelian Inheritance in Man
OS	Overall Survival
p53	Tumour Suppressor Protein 53
PB MNCs	Peripheral Blood Mononuclear Cells
PBS	Phosphate Buffered Saline
PCLO	Presynaptic Cyto-matrix Protein
PCR	Polymerase Chain Reaction
PEG	Polyethylene Glycol
PEST	Proline-Glutamate-Serine and threonine
pg	Picogram
PGM	Personal Genomic Machine
PI3K	Phosphatidyl-Inositol-3 Kinase
PIM1	Pim-1 proto-oncogene, Serine-Threonine-Kinase
PLCγ2	Phospholipase Cγ2
PLEKHG5	Pleckstrin Homology And Rho GEF Domain Containing G5
pM	Pico Molar
PolyPhen 2	Polymorphisim Phenotyping
POT1	Protection of Telomere 1
PR	Partial Response
Pro-B -Cells	Progenitor B- lymphocytes
Q20	Phred Score 20
R	Purine
RB1	Retino-Balstoma-1 Gene
RE	Restriction Enzyme Buffer
Ref	Reference
RFLP	Restriction Fragment Length polymorphism
RING	Ring Finger Domain

RIPK1	Receptor Interacting Serine/ Threonine Kinase 1
RNA	Ribonucleic Acid
rpm	Rounds Per Minute
RPS15	Ribosomal Protein S 15
Rv	Reverse
SAM	Sterile Alpha Motif
SAM	Sequence Alignment Map
SAMHD1	SAM Domain and HD Domain 1
SD	Standard Deviation
SE	Standard Error
Sec.	Seconds
SF3B1	Splicing Factor 3 B Subunit 1
SHM	Somatic Hypermutation
SIFT	Scale Invariant Feature Transform (sorting intolerant from tolerant)
sIg	Surface Immunoglobulin
SMARCA2	SWI/SNF Matrix Associated Regulator of Chromatin A2
SMZ	Splenic Marginal Zone
SNP	Single Nucleotide Polymorphism
snRNA	Small Nuclear Ribonucleic Proteins
SOP	Standard Operating Procedure
SSAHA	Sequence Search and Alignment by Hashing Algorithm
SSC	Saline Sodium Citrate
SSCP	Single Strand Conformation Polymorphism
STK	Serum deoxy-Thymidine Kinase
T	Thymine
TAD	Transactivation domain
TAE	Tris-Acetate
TBE	Tris-Borate
Tbp	Tera Base Pair
TFS	Treatment-Free Survival
Ti	Transition

TLR	Toll-Like Receptor Domain
TLR4	Toll Like Receptor 4
TMAP	Torrent Mapping Alignment Programme
TNFAIP3	TNF Alpha Induced Protein 3
Tv	Transversion
TVC	Torrent Variant Caller
Tx	Therapy
U	Uracil
UBA	Ubiquitin Associated Domain
UK	United Kingdom
UM-IGHV	Unmutated IGHV
UNG	Uracil DNA Glycosylase Enzymes
UPD	Uni-Parental Disomy
USCS	The University of California, Santa Cruz genome database
UV	Ultraviolet
V	Voltage
v	Version
VAF	Variant Allele Frequency
Var	Variant
VCF	Variant Call Format
VDJ	Variable-Diversity-Joining
VQ	Variant Quality Score
W	Thymine or Adenine
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WRN	Werner Syndrome Rec Q Like Helicase
Wt	Wild Type
XPO1	Exportin 1
Y	Pyrimidine
ZAP-70	Zeta-Chain-Associated Protein Kinase 70
ZMYM3	Zinc Finger MYM-Type Containing 3

Chapter 1. General introduction

1.1. Chronic lymphocytic leukaemia

Chronic Lymphocytic Leukaemia (CLL) is a malignant disorder of mature B-lymphocytes. These monoclonal B-cells accumulate in the blood, bone marrow and lymphoid tissues.

1.1.1. Epidemiology

CLL is the most common leukaemia in adults in the Western countries. It accounts for 38% of all leukaemia in the United Kingdom (UK) [1]. It is estimated that 4 - 6 cases are diagnosed in every 100 000 individuals each year. The disease is more common in men than women with a male to female ratio of 1.7:1 [2]. It more likely affects the elderly, with the median age at diagnosis being over 70 years [3]. Conversely, only about 1 in 10 CLL patients are reported to be younger than 55 years [4] and it is very rare in children. Having a family history of CLL in first degree relatives is considered to be a risk factor for developing CLL and this risk is two to four folds higher compared to general population [5].

1.1.2. B-lymphocyte maturation and differentiation

Progenitor B-lymphocytes (pro-B cells) originate from haematopoietic stem cells. In bone marrow they initially undergo multistage development to acquire genomic and protein markers for differentiation with the support of micro-environment, including bone marrow stromal cells. Functional re-arrangement of immunoglobulin (Ig) loci occur in an error-prone process which involves combinatorial rearrangement of heavy (H) chain gene segments V (variable) D (diversity) and J (joining) and the light (L) chain segments V and J [6]. This is followed by combination with a constant region μ gene. After translation of the rearranged Ig genes, combination between the surrogate L chain and μ-H chain of IgM forms a B-cell receptor (BCR) in the more mature pre-B cell (pre-BCR) [7]. Expression of the recombined

pre-BCR is important for B-cell survival and further development, since any defective recombination serves as a control point that renders the cells vulnerable to loss through negative selection aided by bone marrow macrophages. Although the capacity of pre-BCR for induction of proliferation signals by binding with ligands is still not fully understood, the surviving pre-B cells are capable to migrate from bone marrow to secondary lymphoid tissues/organs, e.g. lymph nodes and spleen, for further maturation and differentiation [8].

Once migrating to the lymphoid tissues, these so called naïve B-cells, express IgM and IgD (δ-heavy chain) BCR and stay in resting G0 phase. Upon encountering antigen, the B-cells are activated to undergo proliferation and diversity generation in a process called somatic hypermutation of the V regions (SHM). This process generates a wide range of diverse antibodies with high affinity binding sites to various antigenic epitopes [9]. It is mediated by antigen presenting cells and T-helper lymphocytes [10]. It is also in these B-cells in the secondary lymph tissues/organs that class switch recombination (CSR) occur. The CRS is a process mediated by intra-chromosomal recombination event in which alternative heavy chain constant region genes such as α, γ or ε is selected to form IgA, IgG or IgE, respectively. Both SHM and CSR processes are mediated by Activation Induced Cytidine De-aminase (AID) enzymes [11].

Further differentiation of these antigen exposed mature B-cells that are primed for high affinity, results in the production of memory B-cells. They harbour surface immunoglobulins and await specific antigen re-exposure (secondary immune response). In addition, other B-cells secrete their antibodies and terminally differentiate into plasma cells in the bone marrow [12].

1.1.3. CLL cell origin

The definitive cellular origin of CLL is not clear [13, 14], although there are several suggestions based on surface markers and genomic profiles. Initially, it was proposed that CLL arise from naïve $CD5^+$ B-cells. However, later evidence of existence of some similarities between splenic marginal zone (SMZ) B-cells and CLL cells, suggested that SMZ B-cells are the cells from which CLL develops [15]. More recently, following the discovery that clones of CLL have either mutated or unmutated Immunoglobulin Variable Heavy Chain (*IGHV*) genes

[16], it was postulated that CLL cells can arise from two models of cells which are different in their stage of differentiation and history of encountering antigen.

IGHV mutated CLL group are $CD5^+$and $CD27^+$ and their CLL cells represent the mature post germinal centre cells. *IGHV* unmutated CLL originate from mature $CD5^+$ and $CD27^-$ pre germinal centre B-cells which have retained the capacity to proliferate through competent BCR which enhances tumour cell growth and survival [17]. This biological feature possibly explains the clinical aggressiveness of the disease among this group of patients [18]. Furthermore, the abundance of poly reactive or auto reactive BCR among *IGHV* unmutated cases has been reported. For example, a more intense auto-antigen reactivity was profiled in patients with unmutated *IGHV* and Richter transformation compared to mutated *IGHV* patients who have less aggressive disease [19]. Moreover, the preponderance of ZAP-70 and CD38 which are markers of B-cell activation through BCR signalling in this group of patients provided further support to this view [20].

1.1.4. Diagnosis

Diagnosis of CLL is usually made by blood lymphocyte count, morphology and immunophenotype as described in the International Workshop on Chronic Lymphocytic Leukaemia (IWCLL) 2008 guideline [21]. The typical CLL cells in a blood smear are small in size. They have a narrow cytoplasmic rim and a dense nucleus with indistinguishable nucleoli. Minimum 5 x 10^9 clonal B-lymphocytes per litre of blood confirmed by flow cytometry are required to diagnose the disease. Flow cytometry is also used to identify the surface protein markers of CLL cells, which is characterised by co expression of a T-cell antigen CD5 and B-cell antigens CD19, CD23, and low level of surface immunoglobulin (sIg), CD20 and CD79b [21].

1.1.5. Clinical staging

For the purpose of clinical management of this disease, two clinical staging systems were developed over thirty five years ago by Rai and Binet. The recently modified Rai staging system defines three risk categories. Low risk disease is characterised by lymphocytosis. If

lymphocytosis and lymph node enlargement or organomegaly coexist, then the disease is classified as intermediate risk. If features of bone marrow failure such as anaemia or thrombocytopenia develop, then a highly risky disease is defined [21].

Binet staging system uses similar parameters; the number of involved anatomical lymph node areas is taken into account. Stage A is defined when ≤ 2 lymph node areas involved without anaemia or thrombocytopenia. If > 2 lymph node areas involved, then the disease is classified as stage B. If anaemia and /or thrombocytopenia develop, then the disease will be defined as stage C, irrespective of the presence of organomegaly [21].

These staging systems are still simple and useful for guidance for initiation of therapy in CLL. This is because they evaluate disease stages by both physical examination and routine laboratory tests without the need for any expensive or invasive procedure [22]. However, on the other hand, these systems may be not able to sensitively and specifically predict disease progression and relapse due to the lack of information about cellular and molecular pathology of CLL. Therefore, they are not helpful in tailoring patient specific therapy [23].

1.1.6. Clinical course

1.1.6.1. Heterogeneity of disease outcome

Clinical presentation, course and outcome of the disease are very variable among patients. Thus, some patients have an indolent disease which may even be asymptomatic at time of diagnosis, while others may have features of advanced disease such as anaemia, infection, lymph node enlargement and organomegaly [24]. Accordingly, patients with indolent disease have a life span similar to age and sex matched controls, while others suffer from aggressive, therapy resistant disease or transform to a more aggressive disease (Richter transformation) and die within few years of diagnosis [22].

1.1.6.2. Definitions of disease progression and resistance to therapies

As this project is planned to study CLL patients with progressive and /or therapy-resistant disease, it is important to mention the criteria for defining these disease entities according to the updated guidelines in 2008 for diagnosis and treatment of CLL by the IWCLL [21].

Progressive disease before therapy is defined when one or more of the following criteria are met:

1. progressive increase in the number of circulating lymphocytes of more than 50% within a period of two months or lymphocyte doubling in less than six months
2. development of features of marrow failure
3. symptomatic progressive splenomegaly ≥ 6 cm below the costal margin or lymphadenopathy ≥ 10 cm in longest diameter
4. auto immune cytopenias that is resistant to standard therapy
5. appearance of constitutional symptoms attributable to CLL such as unexplained weight loss of 10% within six months or easy fatigability that renders the patient unable to perform usual daily activities or unexplained fever (body temperature > 38 ◦C) for two weeks or nocturnal sweating for longer than one month without infection [21].

Two types of resistance are described at 2008 IWCLL; refractory disease and relapse. The former is defined as failure to achieve either complete or partial remission after treatment, or disease progression within 6 months of treatment. The latter is defined when disease progression occurred within or after six months of achieved initial response to treatment (complete or partial remission) [21].

Disease progression during or after treatment is characterised by development of at least one of the followings:

1. new lymph node or organ infiltration or an increase by 50% or more in largest diameter of any previously determined site
2. increase in blood lymphocyte count by 50% or more with at least 5000 B-lymphocytes/µl
3. transformation to a more aggressive histology established by lymph node biopsy
4. occurrence of disease related cytopenias after exclusion of drug related cytopenias

1.1.7. Treatment

As mentioned earlier, starting therapy depends on clinical stage and status of disease progression. Patients at early stage or with an asymptomatic and non-progressive disease should usually be monitored without therapy. This is largely because the disease is incurable and there is no evidence to support that early treatment can prolong survival in such patients [25]. In addition, the side effects of unnecessary chemotherapy can increase risks to patients including secondary tumours [21]. For these reasons, treatment is usually confined to progressive or symptomatic disease, for example those at intermediate or high risk group in Rai system or at Binet Stage B and C.

1.1.7.1. Front-line therapies

Before the invention of purine analogues, alkylating agents such as cyclophosphamide and chlorambucil were used to treat the patients. Other alkylating agents, such as bendamustine has also recently been incorporated into first line therapy regimens. Although each drug in this group has its own mechanism of action, they all alkylate DNA, RNA and protein of tumour cells. The primary cytotoxic effect is considered to be caused by formation of DNA cross links and activation of apoptosis [26]. These agents have been studied in clinical trials both as mono-agent and as combination therapy with the goal of improving survival and maximising the attainment of complete response. For example, in the UK CLL 4 trial, responses to treatment in groups of patients randomly assigned to different therapeutic regimens such as fludarabine plus cyclophosphamide or fludarabine alone were compared. A significant improvement in complete response rate was achieved for fludarabine plus cyclophosphamide than for fludarabine alone [27]. Moreover, in German CLL 11 trial, it has been shown that the combination of an anti-CD20 antibody with chlorambucil prolongs remission and improves the overall survival in previously untreated patients with CLL compared to chlorambucil alone [28]. Despite this improvement in treatment response offered by multi-agent therapy and achievement of initial complete response, majority of patients eventually develop resistance.

General mechanisms of chemotherapy resistance including changes in drug transport and metabolism, as well as alteration in apoptotic and DNA repair activity of the tumour cells have been proposed [26]. In CLL, it has been suggested that the first two mechanisms are the least likely causes of resistance to alkylating agents. This suggestion is based on results that found similar pattern of drug uptake between B-cells taken from untreated and resistant CLL patients. Moreover, there are conflicting reports about drug metabolism and detoxification mechanisms being predominant in the cause of resistance to these agents [26]. The main pathway that has been implicated in CLL drug resistance is alteration of DNA repair and apoptotic pathways [29, 30].

Following the alkylating agents, purine analogues such as fludarabine, cladribine and pentostatin have been used in CLL treatment. However, it has been shown that their mechanism of action is different between resting and proliferating cells. Combination of these drugs with the alkylating agents has synergistic action and is associated with longer overall survival in CLL, but with increased toxic side effects [27, 31].

Both the alkylating agents and purine analogues initiate killing of CLL cell by DNA damage [32]. This induces apoptosis via activation of p53 [33]. Therefore defects in either p53 or its upstream activator ATM are associated with resistance to those chemotherapies [32].

More recently, monoclonal antibodies targeting B-cell specific antigen CD20 are introduced in the treatment of CLL. These antibodies include initially used rituximab and more recently available ofatumumab [34] and obinutuzumab (GA-101) [35]. The latter two antibodies have different orientation and more stable binding to their targets.

Unlike chemotherapeutic agents, these antibodies act by complement-dependent cell cytotoxicity (CDC) and antibody-dependent cellular cytotoxicity (ADCC). In a trial conducted by Laurent et al. for relapsed/refractory CLL, the efficacy of ofatumumab in treatment of patients with or without *TP53* mutations was similar [36].

1.1.7.2. Second line therapies

For those high-risk patients, second line treatment has been developed to overcome drug resistance in CLL. For example, alemtuzumab is an anti CD52 antibody that induces cell death through pathways that is p53 independent. Therefore, a combination of high dose steroids and alemtuzumab had been preferably used in patients with fludarabine resistance and defective p53 function [25, 37]. Since alemtuzumab has been discontinued from the market, a later study used Lenalidomide which is a thalidomide analogue that acts through immune modulation by blocking the induction of cancer cell-induced T-cell tolerance and by inhibition of angiogenesis. In that study, patients who had been treated with alemtuzumab and dexamethasone were randomly assigned to no maintenance therapy or maintenance therapy using lenalidomide. The latter regimen resulted in longer progression-free survival and therefore enhanced efficacy of alemtuzumab and dexamethasone [38]. However, there is controversy regarding the use of this agent as a frontline therapy owing to the suboptimal response rate it achieved [25].

Moreover, pathway specific therapies are in continuous development. Idelalisib (CAL101) is an inhibitor of PI3K signalling. This signalling pathway offers protective function for the CLL cells with the support of micro-environment through CD40 ligand [29]. PI3K is expressed at high levels in CLL cells. CAL101 is a high affinity binder of PI3K-α, its cytotoxic activity is partially dependent on caspase activation, but largely independent on p53 [39]. It has been shown that combination of idelalisib and rituximab in relapsed/refractory CLL patients with unfavourable genetic features such as 17p deletion or *TP53* mutations or unmutated *IGHV*, was strongly associated with longer progression-free survival compared to rituximab monotherapy [40].

Another example is ibrutinib, which irreversibly inhibits bruton a gamma-globulinemia tyrosine kinase (BTK). It inhibits BCR signalling which is responsible for B-cell proliferation and it also inhibits cellular adhesion and signals for proliferation from the supportive micro-environment. It has been used in clinical trials for treatment of therapy-naive or relapsed CLL patients with aberrant p53 function [29]. The superior efficacy of ibrutinib as a mono-agent therapy has been shown compared to ofatumumab in unfit patients with relapsed or

refractory CLL [41]. Resistance to this agent has recently been shown to be associated with acquired mutations in *BTK* and in a gene downstream to BTK that is phospholipase Cγ2 (*PLCG2*) detected by NGS [42, 43].

In addition, a selective anti-apoptotic protein BCL2 inhibitor ABT-199 (Venetoclax) has been tested in clinical trials. Its effectivity is mediated by mimicking pro-apoptotic BH3-family members against the over expressed anti-apoptotic proteins in the BCL2 family. This inhibitor is particularly useful for patients with aberrant p53 (upstream negative regulator of BCL2) and with chemoresistance [44]. It has achieved promising results in the treatment of patients with relapse and refractory disease [45].

More recently, selinexor which is a selective inhibitor of exportin1 (XPO1) has been developed. *XPO1* is recurrently mutated in CLL (it is one of the target genes in this thesis) and XPO1 protein is overexpressed in primary CLL cells compared to normal control B-cells [46]. It has been demonstrated that inhibition of XPO1 by selective inhibitors of nuclear export can induce killing of CLL cells in vitro [46].

1.2. General mutational landscape and concepts of carcinogenesis

In the initial attempts to understand the causes of cancer, hereditary influences were identified. This is due to the fact that higher incidence of cancer is reported among patients with hereditary genetic diseases for example ataxia telangiectasia and high incidence of lymphoproliferative disorders [322]. For instance tumour viruses are involved in the causation of malignant disorders such as Kaposi sarcoma and association with human immune-deficiency viruses. These and other acquired mutagenic factors such as irradiation, chemotherapy or any mutation that occur during the lifetime of an individual form the basis of somatic mutation evolution at cellular level. A small proportion of these mutations that confer cell growth and survival advantages will expand clonally to form a tumour. A wide variety of gene families have been extensively studied in different types of cancer for example lung, breast and colorectal cancers. Substantial variation in the number of mutated genes per tumour has been identified. Moreover, within the tumour subtypes, few

commonly mutated genes were found. This observation re-enforced the idea of random evolution and heterogeneity within a tumour subtype. For example, MAPK pathway in lung cancer is most commonly affected pathway, however the incidence of individual gene mutations within this pathway is less than 1% [323]. Similarly in CLL, mutations in various signalling pathways have been implicated in disease heterogeneity which was originally considered as a homogenous disease. Therefore, prognostic biomarkers are continuously developed.

For a long time, conventional clinical and laboratory findings including that by physical examination, patterns of marrow infiltration, lymphocyte doubling time and serum levels of lactate dehydrogenase (LDH) and β2-Microglobulin (β2-M) [47], and the clinical staging systems were used to predict the outcome. For example, LDH and serum deoxy-thymidine kinase (STK) are metabolic enzymes released from proliferating tumour cells. Elevated levels of these enzymes were found in the serum taken from CLL patients. It is widely believed that they reflect the tumour burden and their potential as an independent prognostic marker has been shown [48]. For example, a positive correlation between levels of serum β2-M and clinical stage was also established. Patients with bulky disease as well as those with diffuse marrow infiltration had remarkably higher levels of β2-M compared to others [49]. Other biomarkers such as adhesion molecules and inflammatory mediators such as interleukins and soluble CD23 [50] have been considered and linked with bulky disease as well [51]. However these conventional markers do not fully explain biological causes of the clinical heterogeneity in CLL. Decades passed with no breakthrough achieved. Later, cytogenetic studies revealed chromosomal aberrations and their associations with each subtype (detailed in section 1.2.3). Thereafter, in the late 1990s two biologically distinct subtypes of CLL were discovered based on *IGHV* mutational status [18, 52]. Subsequently, larger studies were performed to study gene expression profile of each subtype either with mutated *IGHV* or with unmutated *IGHV*. This led to the discovery of genes that were differentially expressed between the two subgroups such as ZAP70 and CD38 that overexpressed in the *IGHV* unmutated subtype. These new prognostic parameters have obviously improved the risk stratification for disease progression and response to therapy.

Recently, particular attention has been paid not only to heterogeneity in clinical course, but also to response to treatment. Heterogeneous treatment response has been linked to the existence of underlying biological factors, for example chromosomal aberrations and gene mutational status [53]. Some of these factors have been extensively studied and are now routinely used as established biomarkers for predicting prognosis. Considering the choice of therapy, adjustment of therapeutic decision is made if *TP53* mutations or chromosome 17p deletion exist, so that DNA damaging agents are avoided [54].

As more biomarkers are available, they have been categorised in to three main types: molecular, immune-phenotypic and genomic markers [47]. The most commonly used biomarkers and their roles in biological and clinical stratification of CLL are presented in the following sections.

1.2.1. *IGHV* mutational status as molecular markers

Two different biological subtypes characterising diverse clinical courses have been categorised based on stages of differentiation of the original B-cells. As a marker of differentiation, somatic hypermutation of immuno-globulin variable region is delineated. This process occurs in the germinal centre, thus pre-germinal centre B-cells are less mature and they express unmutated *IGHV*.

Sequence homology of the heavy chain variable region gene to the appropriate germ-line sequences of 98% or more is accounted as unmutated. This cut off was selected to make clinically relevant distinction between the two subtypes as up to 2% disparity can exist due to occurrences of common polymorphisms in this region [55]. The somatic hyper-mutations can occur in the light chains variable region, but at lower rates [56]. Usually they carry two to three times less mutation load compared to mutation load of the variable heavy chain [57]. However, initially the sequence of the light chain was proposed to be used for assessment of the mutation status [18], but later studies revealed that only *IGHV* mutation status at diagnosis is independent prognostic marker [58].

It has been documented that up to 50% of CLL patients are *IGHV* unmutated [16]. As it arises from naive B-cells, it tends to represent a more aggressive disease [18]. Associations between this subtype and a typical CLL cell morphology, trisomy 12 [59], higher percentage of CD38 and higher expression of ZAP-70 [60] have been documented. These patients commonly suffer from rapid disease progression, poor response to multi-regimen chemotherapy and shorter survival [52]. A significant tendency for V1 family gene usage was found, especially biased usage of V1-69 among this group and this is also associated with an unfavourable outcome [61, 62].

IGHV mutated CLL arises from memory B-cells. It has a relatively indolent course and better prognosis compared to patients with unmutated *IGHV* [18]. Moreover, association of this subtype with del 13q14 has also been found [52]. Notably, preferential usage of V3-21 repertoire among this subtype was associated with p53 dysfunction [63, 64] and shorter overall survival than other mutated IGV repertoires [65].

1.2.2. Immuno- phenotypic features

1.2.2.1. CD38

CD38 is a trans-membrane glycoprotein, its expression occurs at times during B-cell maturation when interactions between the cells are necessary for development [66]. Surface expression of CD38 is based on a cut-off value of 30% determined by flow cytometry [67]. CD38 expression is considered as a robust biomarker to predict clinical outcome in the majority of patients owing to its stability over time of disease progression or upon treatment with chemotherapy [68]. It has been suggested that two subsets of CLL can be distinguished based on CD38 expression [17].

CD38 negative patients are characterised by longer treatment-free time and prolonged survival [67]. In contrast, CD38 positivity is associated with unfavourable clinical course including a more advanced disease stage [69]. Although initially a strong association with

unmutated *IGHV* was found [70], but later studies found that the association was not absolute and CD38 should rather be used as an independent prognostic indicator [71].

1.2.2.2. ZAP-70

ZAP-70 is a CD3 associated protein kinase of T cells. ZAP-70 activation triggers its downstream signalling pathways including the Ras/mitogen activated protein kinase (MAPK) pathway which is important for T-cell survival and proliferation [72]. Typically B-cells lack ZAP-70. They utilize a different protein kinase which is Syk to mediate signal transduction of the BCR complex. Thus, ZAP-70 and Syk are alike in their roles in membrane antigen-receptor signalling pathways [73].

In CLL, cells with unmutated *IGHV* genes commonly express high levels of ZAP-70 protein while CLL cells with mutated *IGHV* frequently associate with undetectable levels of ZAP-70 protein [73, 74]. Moreover, it has been shown that high-level expression of CD38 is correlated with elevated ZAP-70 expression [73].

In addition of protein expression, later studies have suggested that ZAP-70 mRNA expression might serve as a surrogate marker for identifying the Ig-mutated and Ig-unmutated CLL subtypes and prognostic subgroups. For example, patients whose CLL cells express a high level of ZAP-70 usually have progressive disease with shorter overall survival compared to cases with low level of ZAP-70 mRNA expression [60].

1.2.3. Genomic aberrations

Genomic aberrations in CLL vary considerably from patient to patient. Initially, by using conventional G banding, 40 - 50% of the patients had been reported to have chromosomal abnormalities. However, the low mitotic activity of the CLL cells in vitro hampered such analysis. Later, with the invention of fluorescence in situ hybridisation (FISH), the frequency of detected clonal chromosomal abnormalities was found to be higher and reached up to 80%. This was due to its capability to detect such genetic lesions (copy number changes) in

interphase nuclei [75]. The most common recurrent aberrations in CLL are deletion (del) 13q, del 11q, trisomy 12q and del 17p [75].

Copy number neutral loss of heterozygosity (CNN- LOH) or uni-parental disomy (UPD) (imbalance of specific allele without net change in its copy number) occurs because of mitotic recombination or gene conversion and this is not uncommon in CLL [76]. Such abnormalities can be detected by high density single nucleotide polymorphism array (SNP array) [77]. The use of genome arrays has led to the identification of additional genetic lesions which were otherwise undetected with FISH, for example; trisomy 19 [78], gain of 2p [79] and identification of anatomical subtypes of del 13q14 in CLL [80].

All in all, these changes have been used as prognostic indicators. Some of them are linked with indolent disease while others are associated with rapid disease progression and chemoresistance. The most common and intensively investigated chromosomal abnormalities are discussed individually in subsequent sections.

1.2.3.1. Structural chromosomal abnormalities

1.2.3.1.1. Del.13q

Deletion of the long arm of chromosome 13 involving 13q14 is the most frequent chromosomal abnormality observed in CLL. Incidence approaching 55% has been reported in unselected CLL cohorts [81]. It can either occur as a single genetic abnormality or in association with other abnormalities. The size of the deletion varies from the entire loss of the long arm to a few megabases which can be detected by FISH. The majority of the deletions are monoallelic, but biallelic deletions do occur [82].

Initially, it was proposed that existence of variation in clinical outcome in patients with this lesion as a solitary abnormality may be due to presence of differences either in the number of lost alleles or the sum of interphase nuclei that carry the lesion. For example, it has been documented that patients with monoallelic loss carry a better prognosis than those with biallelic loss [82]. However, subsequent studies found that the nuclear percentage carrying the abnormality affects the prognosis rather than the number of affected alleles [83]. Later,

the size and anatomical location of the affected regions were also identified to cause this variation. The minimal deleted region (MDR) which contains target genes has been defined. In 13q deletion, this region encompasses sequences encoding deleted in leukaemia (*DLEU-2*) and (*DLEU-1*) [84, 85]. There are also two micro RNAs (small non- coding RNAs that are 21-22 nucleotide long and contribute to regulation of coding genes' expression) miR-15a and miR-16-1 which are located in the MDR [86]. These micro RNAs are found to share tumour suppression activity by their interaction with pathways controlling apoptosis. An example is direct negative regulation of p53 [87]. This is further supported by a finding that loss of miR-15a and miR16-1 is associated with de-repression of BCL2 oncogene and an adverse outcome in CLL [88].

It is still unclear why not all the patients with deletion 13q14 have unfavourable prognosis. To refine the prognostic influence of this genetic lesion, it has been proposed to classify del13q into two groups; group I, where the deletion involves a narrower region, it is confined to MDR and is associated with a relatively indolent disease [89]; whereas in group II, deletions involve a larger area that include the tumour suppressor retinoblastoma-1 (*RB1*) gene. This group is more commonly associated with aggressive clinical course and transformation to diffuse large B- cell lymphoma (DLBCL) [80, 90].

More recently, other genes outside the region of the MDR were found to be involved and correlated with adverse prognostic outcome [91]. In addition, a study has found an association of 13q deletion with *MYD88* mutation [92] that is a driver event in CLL [93].

1.2.3.1.2. Del.11q

Initially, the frequency of this abnormality in CLL was underestimated when detected by conventional chromosomal banding. Later studies using FISH for screening of CLL cases revealed the true prevalence of this chromosomal aberration. By far, this genetic aberration is ranked as the second most common chromosomal lesion which occurs in up to 10% of CLL patients at time of diagnosis [94, 95].

The ataxia-telangiectasia mutated (*ATM*) gene maps to chromosome 11q22-23 [96]. It is one of the candidate genes within the MDR affected by this karyotypic abnormality in CLL [97].

This gene is a member of the phosphatidylinositol-3 kinase (PI3K) family [95]. Its down regulation has been observed in patients with deletion of long arm of chromosome 11 carried by > 20% of the nuclei as detected by FISH [95]. Similarly, sub-chromosomal alterations were found to have the same consequences [98].

Del11q can be associated with mutation of the remaining *ATM* allele in 36% of CLL cases. In this respect, two subgroups based on the integrity of the residual *ATM* allele have been identified; firstly, patients with monoallelic *ATM* loss and; secondly, patients with biallelic *ATM* defects and complete loss of *ATM* function. The latter group were characterised by impaired responses to cytotoxic agent in vitro and worse prognosis [99]. It has been suggested that subclones carrying deleted and/or mutant *ATM* can develop during disease progression and lead to selection and further expansion of these subclones [99]. Furthermore, 11q deletion is considered to be the most common karyotypic abnormality to undergo clonal expansion with disease progression [100]. Correlation study revealed that del11q is significantly associated with poor prognostic features including extensive nodal involvement and more rapid disease progression in CLL [94].

Another gene that is affected by this deletion is *BIRC3*. 87% of 11q deleted cases harboured allelic loss of this gene, with only a small percentage of patients having a mutation on the residual allele (< 5%) [101]. However, a later study found a more frequent co-occurrence of *BIRC3* mutation and loss of the other allele which occurred in a higher percentage than originally thought [102]. Deletion or mutation of this gene was found to be associated with drug resistance in *TP53* unmutated cases [102]. However, a more recent study showed that larger 11q deletion that involves *BIRC3* does not differ from 11q deletions not involving *BIRC3* in terms of impact on survival [103]. The same study also reported higher association of patients with *BIRC3* mutation and/or deletion with other poor prognostic factors compared to patients without these abnormalities [103].

1.2.3.1.3. Del.17p

Deletion of the short arm of chromosome 17 including the *TP53* locus at 17p13.1 are found to occur in up to 7% of treatment naïve patients [104] and in up to 50% of relapsed or refractory cases [105]. This deletion usually associates with *TP53* mutation on the other allele in up to 80% of patients [106]. Del17p therefore results in loss of the tumour suppressor *TP53* gene and renders CLL cells resistant to spontaneous apoptosis and DNA damaging therapy-induced apoptosis [107].

More recently, *TP53* mutation as a solitary abnormality has been shown to have the same biological and clinical consequences as 17p deletion [108]. 17p deletion with or without mutation in the remaining *TP53* allele has been associated with poor prognostic features [102] such as faster disease progression and poor response to chemotherapy [109]. For that reason, chromosome 17p13 deletion is regarded as the first (hitherto) biological marker to affect decision of choosing the first-line treatment [110]. Patients with this abnormality are offered alternative therapies that are independent of p53 pathway such as monoclonal antibodies directed against epitopes expressed on the surface of tumour cells [111].

1.2.3.2. Numerical chromosomal aberrations

1.2.3.2.1. Trisomy 12

Similar to other chromosomal abnormalities, the conventional G banding analysis underestimated the frequency of this genomic abnormality in CLL. Trisomy 12 is the third most common karyotypic abnormality and the most frequent numerical chromosomal abnormality in CLL. It has been found in 10 - 20% of CLL cases at time of diagnosis [98]. For risk stratification, it has been more commonly recognised as a marker of intermediate risk disease [75]. However, recent studies have considered this aberration as an indicator of low risk disease [112]. This difference is assumed to be due to a combination of associated genetic abnormalities that are located outside chromosome 12 such as recurrent mutations of *NOTCH1* on chromosome 9 which may confer other pathogenic mechanisms [113, 114].

However, the precise pathogenic mechanisms of how trisomy 12 can cause tumorigenic transformation are still unclear because multiple genes are distributed throughout the entire length of the chromosome. Gene expression profiling showed differential expression of

various genes located on this chromosome. One of the most important genes is murine double minute 2 (*MDM2*) which is involved in the repression of tumour protein p53 [115]. Overexpression of this gene results in de-regulation of cell cycle [116]. More importantly, this genetic lesion is primarily considered as a clonal driver that occurs early in CLL evolution and facilitates the acquisition of subsequent chromosomal aberrations or mutations in genes such as *TP53*, *NOTCH1* and *FBXW7* [117]. Correlations with atypical cellular morphology, high CD38 and ZAP-70 expression as well as unmutated *IGHV* gene have also been reported [118].

1.2.3.2.2. Trisomy 2P

Recently, by application of high density SNP array and high-resolution genomic profiling of a large cohort of unselected CLL cases, this lesion was found in up to 7% of patients with CLL [76]. It was found that the minimally gained region sized approximately 2 Mb and was located at 2p16.1-2p15 which includes multiple genes including *BCLIIA* and *XPO1* [119]. Patients with this karyotypic lesion carried high-risk genomic changes.

1.2.3.3. Complex karyotype

Karyotype complexity is defined by the existence of three or more cytogenetic abnormalities in a single case. It has been found in up to 16% of CLL cases [120]. Carrying this abnormality has been stratified among high risk disease as it predicts rapid disease progression, short duration of remission and shorter overall survival [121]. The mechanism underlying this complexity is thought to be due to genomic instability and has been linked to ATM/p53 aberrations [122].

1.2.3.4. Chromosomal translocations

Unlike the other neoplasms, the presence of specific chromosomal translocations is not a characteristic cytogenetic feature of CLL. However, an incidence of 32 - 42% has been reported in various studies of CLL [123]. It has been found that these chromosomal rearrangements have significant prognostic potential. Unbalanced translocations have been linked to poor outcome regardless of the sum of the rearrangements [124]. Translocations

involving IGH locus on chromosome 14 are not common in CLL, although a disease subset with poor outcome was identified in association with this translocation [125].
In subsequent studies, it has been revealed that the outcome of the translocation (t) is determined by the recipient chromosome involved in the partnership. Thus, t (14; 19) which involves the *BCL3* locus is linked to complex cytogenetics, germ-line *IGHV*, atypical cellular morphology and inferior survival, whereas patients harbouring t (14; 18) that involve *BCL2* locus were not marked by aggressive prognostic features [125].

Other recurrent translocations have been reported, for example translocation involving chromosome 13q which occurs in 10% of patients carrying chromosome 13q loss [126]. Such abnormality is found to be associated with larger 13q deletion that involve the *RB1* locus and hence, can cause poor prognosis [127]. However, being a balanced translocation in most of the cases, such abnormality has not been implicated in causation of poor prognostic features [128].

1.2.3.5. Chromothripsis

A unique pattern of clustered rearrangements typically involving only a single chromosome and presenting at least 10 switches between 2 or 3 copy number states is defined as chromothripsis. It is thought that this chromosomal aberration results from a single cellular catastrophe. This event is not common in CLL and found only in <5% of previously untreated patients [119]. However, higher frequencies have also been found especially in patients with p53 aberrations [129]. Such abnormality is associated with unmutated *IGHV*, rapid disease progression and shorter overall survival [119].

1.2.4. Recurrent gene mutations

As mentioned earlier in this chapter, clinical staging systems do not provide biological reasons for the clinical variability and are not reliably predict disease progression and drug resistance in CLL. Obviously, biomarkers developed on knowledge about pathogenesis of CLL especially that about genomic alterations, are required. To stratify high-risk disease,

cytogenetic studies have correlated del17p and del11q with *TP53* and *ATM* mutations, respectively [99, 105].

1.2.4.1. *TP53*

The *TP53* gene is mapped on the short arm of chromosome 17 (17p13.1). It composed of 11 exons of which exon 1 is non-coding. The gene encodes transcription factor p53 protein which is composed of 393 amino-acid residues. Structurally, p53 protein has 3 main domains. Firstly, the N- terminally located transcription activation domain which spans from amino acid 1 - 91 residues; secondly, a middle sequence specific domain that binds to target DNA, the so called DNA binding domain which extends from amino acid 95 to 288; and thirdly, the C- terminally located nuclear localisation/tetramerisation and basic domain which extends from amino acid 316 to 393 [130].

The activities of p53 including activation of apoptosis and cell cycle arrest, are determined by folding, tetramerisation and nuclear localisation functions which indirectly affect the DNA binding capacity of the protein. As evidenced by various experimental studies, the activation domain induces pro-apoptotic genes or inhibits anti-apoptotic genes. On the other hand, maintenance of the activation domain's trans-activation or trans-repression functions is the duty of the C- terminal domain [130].

In normal cellular state, p53 is present at trace amounts due to its short half-life. Under certain conditions, for example, DNA damage, the protein becomes activated to adapt its shape to interact with the damaged DNA. This is mediated by kinases which induce conformational changes and tetramerisations of p53 to become large enough to be trapped inside the nucleus [130].

P53 defects, either due to allelic loss or mutations, can occur in various solid tumours as well as other haematological malignancies [131]. As mentioned above, the defect is usually caused by its deletion from one allele and mutation in the remaining allele [132]. The mutations are distributed throughout the coding sequences of the gene. However, the functional characteristics of the mutations vary according to the tumour type. Most of the

mutations are missense, so that the full length of the protein can be produced but the change in amino acid hampers its DNA binding ability [133]. The same findings are common in CLL, in that 89.9% of mutations are localised to the DNA binding domain in exons 5 - 8 while the remaining occur in other exons [134].

TP53 mutations are found in 5 - 10% of previously untreated CLL patients [135]. The frequency of mutations tends to increase over disease progression. Thus, the mutations are more likely detected in patients with a history of therapy, with an incidence of 25% has been reported among chemotherapy-resistant patients [106, 136]. Regarding the nucleotide changes caused by the mutation, transitions are common, but incidence of CpG transition compared to other cancers is significantly low. Classical p53 mutation hot spots for CLL are found at codons 175, 179, 220, 248, 273, 281 and 209 [134].

Owing to the *TP53* mutation profile and residual function similarity in patients without and with treatment, it was suggested that the post-treatment mutations are selected, rather than being caused by chemotherapy [134]. This hypothesis was supported later in longitudinal studies using highly sensitive NGS methods. It has been suggested that CLL subclones with *TP53* mutation exist before and expand after treatment induced selection [137]. To further support this, a recent NGS study for 309 CLL patients at early stages has found that up to 9% harboured small subclones with *TP53* mutations. Importantly, these patients shared the same aggressive clinical phenotype and poor survival as those harboured clonal *TP53* aberrations [138].

Although *TP53* mutations are more common in patients with unmutated *IGHV*, they are independently associated with poor disease outcome in CLL, predominantly by reducing apoptosis and inducing resistance to DNA damaging agents [139]. Of notes, the clinical consequence may vary depending on location of mutations and their effects on p53 functions. For example, It has been documented that mutations in DNA binding motif and gain of function mutations are associated with extremely poor survival in CLL [139].

1.2.4.2. *ATM*

This gene is mapped on chromosome 11 (11q22.3), it composed of 63 exons of which exon 1 is non-coding. It encodes the phospho-protein ATM which is composed of 3065 amino-acid residues. It has multiple domains including the C- terminal domain which shares homology with phosphatidylinositol 3-4 kinase (PI3K), phosphokinase like domain, and ARM-TF domain [140]. This phospho-protein acts upstream to p53 in controlling the cell cycle and apoptosis through mediation of p53 phosphorylation at multiple sites in response to DNA damage [141].

Mutations of this gene can occur in up to 12% of unselected CLL patients [142], while monoallelic loss through 11q abnormality occurs in 20%. Moreover, up to 36% of 11q deleted patients harbour mutated *ATM* gene in the remaining allele [99]. A small proportion of *ATM* mutations can occur without loss of *ATM* allele [95]. *ATM* mutation and/or 11q deletion occurs in 25% of cases at time of diagnosis [95]. Mutations are distributed across the entire gene sequence. Although no hot spots have been identified, more frequently the C- terminus of the protein (PI3K domain) which is critical for phosphorylation activity of the protein is affected. Regarding mutation types, those with truncating features accounts for 55% of all the mutations while others are missense mutations [143].
Consequently, biallelic *ATM* loss due to *ATM* deletion and mutations and/or mutations alone diminishes ATM expression [144]. This reduces phosphorylation of target genes and lowers auto-phosphorylation in response to radiation in CLL [33, 99, 145]. Therefore, defects in *ATM* gene can produce effects similar to p53 loss [33].

In a clinical context, *ATM* defects were associated with chemoresistance, advanced disease stages and shorter overall survival [146]. Mutations can occur in both *IGHV* mutated and unmutated groups, but are more commonly associated with germ-line *IGHV*. Occasionally, it can also co-exist with *TP53* mutations [142].

1.2.4.3. Novel recurrent gene mutations

Despite the earlier biomarkers for CLL mentioned above, there have been continuous efforts to identify new biomarkers. This is due to the fact that the available biomarkers failed to

predict the outcome for all the CLL patients. The initial effort using whole genome sequencing (WGS) for 7 cases led to identification of new recurrent gene mutations with unknown clinical significance in CLL. Later, studies using whole exome sequencing (WES) for a substantial number of CLL cases identified somatic mutations in around 75 novel genes. It is estimated that a CLL genome can carry somatic mutations ranging from 2 to 76 mutations. As revealed by the WES studies, most of these mutations were detected in less than 5% of cases.

For those recently identified gene mutations, the exact functional consequences and their role in CLL progression are still not very clear [147]. However, some of these genes have been evaluated in the context of a larger series of patients, in addition of a correlation between higher number of somatic mutated genes and unmutated *IGHV* [147].

Interestingly, the incidences of various gene mutations were different between CLL subtypes. For example, in *IGHV*-unmutated cases, mutations of *NOTCH1, SF3B1, XPO1* and *POT1* were commonly found [148, 149]. In contrast, *IGHV*-mutated patients had enriched mutations of *KLHL6, MYD88* and *CHD2* [148]. A number of important signalling pathways relevant to CLL pathology have been identified through functional clustering analysis of the commonly mutated genes, they include NOTCH1 signalling, mRNA splicing, DNA damage – cell cycle- control and inflammatory response pathways [150]. All of those suggested that such mutations play roles in CLL development.

1.2.4.3.1. *NOTCH1*

This gene is mapped on chromosome 9q34.3. It composed of 34 coding exons. It encodes a transmembrane protein (receptor) which is composed of two main domains; namely an extracellular N-terminal domain and an intracellular C-terminal domain. The latter composed of ankyrin (ANK), transactivation (TAD) and proline-glutamate-serine and threonine (PEST) domains. The protein acts as a ligand activated transcription factor in controlling cellular proliferation and apoptosis [151].

In quiescent cells, the two domains are joined together. Upon ligand binding to the extracellular domain, the intracellular domain releases through proteolytic activity of gamma

secretase. The detached intracellular domain moves to the nucleus to trigger signals upon which active transcription complexes are accumulated. It is the PEST domain part of C-terminal of the protein that is responsible for the limitation of this activity. This function is accomplished through phosphorylation of the PEST domain by which proteasomal degradation via the FBXW7 complex is enhanced [147].

In CLL, almost all of the mutations detected are located in the exon 34, which encodes the TAD and PEST domains. Among the mutations, p.F2482Ffs*2 is the most common alteration, making up more than 85% of mutations in this gene. Usually, as a result of these mutations, a truncated highly stable protein is generated that is then overexpressed in the CLL cells [152]. Clinically, *NOTCH1* mutations characterise patients with more aggressive disease. In addition, this aberration is enriched in patients with unmutated *IGHV* and trisomy 12 [153]. Although the mutations can be detected at time of diagnosis and in non-progressive CLL, few cases showed acquisition or loss of *NOTCH1* mutations over time of disease progression or after receiving treatment [153]. Thus, the incidence of *NOTCH1* mutation varies from 4% in unselected CLL cohorts [154] to 20% in chemo-refractory CLL. An even higher incidence up to 30% was found among CLL patients with Richter transformation [155].

1.2.4.3.2. *SF3B1*

SF3B1 gene is located on chromosome 2q33.1. It composed of 25 coding exons. It encodes SF3B1 protein that is composed of two regions, the N-terminal hydrophilic region which contains a number of protein binding motifs (encoded by codons 1 - 450) and the C-terminal region, which consists of 22 non identical HEAT repeats (codons 453 - 1298). It is a component of the spliceosomal complex, which also contains five small nuclear ribonucleic-proteins (snRNP) [156]. This complex becomes assembled on premature RNA to efficiently splice introns with fidelity. Mutations affecting splicing recognition sites may result in the abnormal splicing of various genes in the form of retained introns, skipped exons, abnormal elongation or truncation of proteins and altered gene expression [156]. A recent study in CLL has documented that mutated *SF3B1* is associated with defective ATM/p53 transcriptional and apoptotic responses to DNA-damaging agents even in samples with intact *ATM* and *TP53* genes [157].

Somatic alterations of *SF3B1* occur in 10 - 15% of CLL patients at time of diagnosis [158] and up to 30% of chemo-refractory cases [159]. 90% of the identified mutations are clustered in exons 14 - 16 that encodes the HEAT 5 - 8 domains. The four most common hot spots at codon 700, 742 (reported for the first time during this study), 662 and 666 together comprise 80% of all the *SF3B1* mutations in CLL [158].

Considering their clinical effects, these mutations more frequently occur in patients with advanced disease who showed shorter time to progression and poorer response to treatment. Furthermore, associations with adverse biological features such as unmutated *IGHV* and 11q abnormality have also been documented [159].

1.2.4.3.3. *BIRC3*

BIRC3 is localised on 11q22.2. Structurally, it consists of 9 exons of which the 1st exon is non-coding. BIRC3 protein has multiple domains including BIR1-3, ubiquitin associated domain (UBA), caspase recruitment (CARD) and Ring finger (RING) domains. Codons 529-604 encode the RING domain which is the key player of unbiqutination activity to regulate NFκB signalling. This gene is a negative regulator of NFκB. Target genes of NFκB are involved in inhibition of apoptosis, signal transduction and chemotaxis. These targets have significantly higher expression in lymph node derived CLL cells compared to CLL cells in the peripheral blood [160], suggesting the existence of specific interactions between protective micro-environmental niches and CLL cells. Since activation of NFκB signalling results in pro-survival signals to the leukaemic clones through the up-regulation of those anti-apoptotic genes [160], such activation through different mechanisms including micro- environmental interactions is an important contributor to disease progression and chemotherapy resistance in CLL [161].

At molecular level, gene alterations along the NFκB pathway also play a role in pathogenesis of CLL through the activation of this pathway [162]. This is true for other B-cell malignancies such as splenic marginal zone lymphoma [163] and diffuse large B-cell lymphoma [164]. *BIRC3* monoallelic loss associated with 11q deletion can result in activation of NFκB

pathway. Moreover, truncating mutations (nonsense and frame shift alterations mostly affecting RING domain) and monoallelic loss can have similar effects.

The incidence of *BIRC3* mutations is 4% at time of diagnosis, 10% at time of disease progression and is up to 24% in fludarabine refractory CLL [162]. It characterises high risk genetic and phenotypic features such as unmutated *IGHV*, fludarabine refractoriness and poor survival in CLL [162]. However, independent significance of *BIRC3* alterations in patients with 11q deletion is uncertain.

1.2.4.3.4. *MYD88*

MYD88 is located on chromosome 3p22.2. It composed of 5 coding exons which encode a protein of 317 amino-acids. This protein molecule is composed of 2 main domains including death domain (DD) (amino-acids 1-150) and toll/interleukin receptor domain (amino-acids 151-317) that is encoded by exon 3, 4 and 5. It is found to be recurrently mutated at a low incidence (2% - 10%) in CLL. It encodes a critical adaptor molecule of the interleukin-1 receptor/toll-like receptor (TLR) signalling pathway. Consistent with this functional feature, *MYD88* mutated CLL show secretion of significantly higher amounts of the interleukin-1 receptor antagonist, interleukin 6, and chemokine ligands upon TLR stimulation compared to wild type group [165]. Activation of TLR leads to induction of NFκB signalling [166]. In addition, the excessive production of these cytokines recruits more macrophages and nurse-like cells by CLL cells, forming a favourable niche to escape apoptosis and death [167].

The activating mutations in this gene have been identified in 3% of unselected CLL cases. Their prognostic significance in CLL is still not clear. An early report demonstrated that those mutations were more common in patients $\leq$ 50 years old, with mutated *IGHV*, low CD38 and ZAP-70 expression and favourable outcome [168]. However evidence from a recent study on a bigger CLL cohort (published in 2015) suggested that such mutations are associated with advanced disease stages (Binet stages B and C) and a shorter time to first time treatment (TTFT) [169].

1.2.4.3.5. *POT1*

The gene is localised on chromosome 7q31.33. It consists of 19 exons of which 15 are coding (exons 5 - 19). In the N- terminal portion of the protein, there are two important domains OB1 and OB2 which are essential for DNA binding activity and protection of telomere (POT) function. In normal cell division process, DNA initially needs to undergo replication. With each round of cell division, the length of telomere that is composed of protein bound sequence of GT rich repeats is lost. If the telomere loss is profound, the cell can no longer undergo DNA replication to divide and therefore loses the capacity to survive [170]. To counteract this DNA loss, the enzyme telomerase adds telomeric DNA sequences only to the linear structure of 3' GT rich sequence overhang. Hence, the telomere length is restored and the cell can undergo further cell cycles [170].

POT1 encodes a member of a protein complex that is called shelterin, which binds to the telomeric sequence of DNA. This binding forms a bridge between single strand DNA (the overhang) and the double strand DNA. The bridging results in folding of the telomere to render it inaccessible to the telomerase enzyme. This process is called end capping [171].

The efficiency of this capping is not only dependent on proper binding of the protein complex to DNA, but also depends on the length of the telomere. The shorter the telomere, the less efficient binding is resulted [171]. Therefore, reduction or loss of the capping capacity indicates higher replicative capacity induced by telomerase due to unprotected chromosomal ends as seen in CLL [172]. Furthermore, it is postulated that loss of this DNA capping ability initiates erroneous DNA damage response culminating with inappropriate end repair through end joining (fusion) of sister chromatids and chromosomal aberrations.

In CLL, *POT1* mutations are found in 3.5 - 5% of unselected cases and up to 9% of *IGHV* un-mutated cases [173]. It has been found that these mutations mainly affect the N- terminal OB1 and OB2 domains. The mutations were heterozygous and more commonly resulted in truncated protein. Dominant negative role of mutations has been confirmed by defective DNA binding in a heterozygous *POT1* mutated cell line. However, POT1 mutated protein could form the complex with the other member of the family [173]. Molecular analysis of *POT1* mutations in CLL showed increased incidence of telomere attrition and chromosomal end fusion with more advanced disease while the incidence was lower in patients prior to

disease progression. Specifically patients with dysfunctional telomeres were found to have large-scale genomic rearrangements enriched at the ends of the chromosomes. Moreover, the recent discovery of frequent *POT1* mutations and associated telomeric and chromosomal abnormalities in CLL strongly supports the model that telomere shortening and fusion contribute to the progression of CLL [174]. Consistently, the significantly higher rates of chromosomal fusions and complex karyotype [173] have been found to be associated with shorter survival in CLL [120].

1.2.4.3.6. *SAMHD1*

This gene which is located on chromosome 20q11.23, encodes a protein that has two important domains, the first is sterile alpha motif (SAM) and the second is histidine-aspartic domain (HD). Normally the gene reduces the intracellular dNTP pool by decomposing this molecule through resection of its three phosphate atoms. Depletion of this element which is required for DNA synthesis can result in reduced cellular proliferation through cell cycle inhibition and apoptosis [175].

Whole genome or whole exome sequencing studies of CLL have recently indicated that this gene is recurrently mutated in CLL. Mutations can occur at a frequency ranging from 3% in untreated CLL to 11% of relapsed and refractory cases [175]. The mutations are distributed throughout the exons and are found to occupy the entire clones reflected by high variant allele frequency. Genome wide array has revealed that 80% of *SAMHD1* mutations were associated with monoallelic loss or copy number loss of heterozygosity. This finding suggested that disruption events of this gene can be a founder to early stage leukemogenesis [176].

SAMHD1 mutated CLL cells invariably had reduced mRNA expression compared to normal B-cells. Furthermore, exposure of cell lines carrying recombinant mutated HD domain to escalated doses of DNA damaging agents, showed resistance to cell death [175]. This study highlighted the possible role of HD domain of this gene in the regulation of cell survival and response to DNA damaging agents [176]. From the clinical perspectives, mutations were found to be associated with enhanced resistance to DNA damaging chemotherapy evidenced by enrichment of these mutations in the relapsed/refractory patients [175]. Still no

significant associations for mutations are found with overall survival rate and transformation to Richter syndrome due to small size of the studied samples.

1.2.4.3.7. *CHD2*

CHD2 is situated on chromosome 15q26.1. It is composed of 38 coding exons which encode a complex protein that has multiple domains including N-terminal chromatin organisation modifier domain, dead-like helicase superfamily domain, a putative DNA binding domain and a C-terminal domain of unknown function. *CHD2* is a regulator of transcription and chromatin remodelling [177, 178].

Somatic mutations in *CHD2* have been identified in 5 % of CLL [117]. Notably, 50% of the mutations are predicted to result in deleterious biological effects through truncation of protein structure either by frame-shift alterations, splice site changes and premature stop codons. The mutations were commonly detected in the functional domains particularly the DNA binding domain (codons 480 - 946), accounting for 33% of all the mutations [177].

Transcriptome analysis has identified significant enrichment for genes belonging to the phosphatidylinositol-4, 5-bisphosphate 3-kinase (PI3K) pathway. Since this pathway plays a remarkable role in the BCR signalling which contributes to survival and proliferation of CLL cells [179, 180], a possible role of this gene in CLL development has been suggested.

Biological parameters of CLL patients with *CHD2* mutations showed frequent *IGHV* mutations, however a statistically significant association between these two biological variables could not be identified. Moreover, there were no difference in the clinical parameters of *IGHV* mutated patients with or without *CHD2* mutations [177]. Detailed clinical description of *CHD2* mutations in CLL is still unknown.

1.2.4.3.8. *XPO1*

This gene is located on chromosome 2p15 and composed of 24 coding exons. It encodes XPO1 (exportin-1) protein which functions in nucleo-cytoplasmic export of numerous proteins including anti-apoptotic and tumour suppressor proteins. In fact, the nuclear export

of the tumour suppressor p53 requires both MDM2 and XPO1. Through E3 ubiquitin ligase activity, MDM2 initiates the nuclear export signal (NES) in p53, resulting in a conformational change in p53 that lead to exposure of its NES domain. Following this conformational change in p53, XPO1 recognises the NES signal of p53 and transfer the protein from the nucleus to the cytoplasm. In this way, XPO1 negatively regulates transcriptional activities of p53 [181]. However, an influence of *XPO1* mutations on p53 regulation function is still unknown.

In CLL, this gene is found to be recurrently mutated, with a frequency approaching 3 - 5% of patients using whole genome and whole exome and targeted deep sequencing in unselected cases or untreated patients. Harbouring *XPO1* mutations was considered as a predictor of poor outcome, however due to small sample size there was lack of reliable statistical evidence to prove this [13, 92].

1.2.4.3.9. *FBXW7*

This gene is localised on chromosome 4q31.3 and composed of 11 coding exons. The protein encoded by F-box and WD40 repeat domain-containing 7 (FBXW7) is an ubiquitin ligase. It has three domains including the dimerisation domain, the F-box domain that is responsible for recruitment of the other constituents of the ubiquitin ligase complex, and the WD40 (codons 400 - 700) repeats that attaches to substrates [182]. FBXW7 has central roles in cell division, growth, and differentiation by down regulating several target proteins, including c-Myc, Notch family proteins and cyclin E [183].

Mutations of this gene can have a dominant negative effect, particularly the hot spot codon 465 (located inside the WD domain) which is important for recognition of the target onco-proteins [184]. Accordingly, dysfunctional FBXW7 can be an alternative pathway for activation of NOTCH signalling and consequent MYC transcription [185]. Notably, like p53 disruption, NOTCH signalling and MYC activation are frequently enriched in CLL cases with Richter transformation [186].

Mutations of *FBXW7* were found in 8% of CLL cases with trisomy 12 [13]. In a study that recruited a large number of untreated CLL cases, 2.5% of the patients carried *FBXW7* mutations. The mutations were found to affect the WD40 domain; moreover, alterations in

codon 465 comprised 30% of the mutations. Heterozygous missense mutations were the most frequent alteration accounting for 91% of the mutations. Variant allele frequencies of one third of the detected mutations were below 10% [92].

1.2.4.3.10. *PCLO*

The gene is located on chromosome 7q11.23-21.3. It constitutes of 25 coding exons that encode a giant protein of 5142 amino-acids. The piccolo presynaptic cyto-matrix protein (*PCLO*) is involved in the organisation of the cytoskeleton and extracellular matrix as well as cell adhesion through calcium sensing [187]. For instance, increased intracellular Ca^{2+} concentration has been shown to be important in cell migration. In human lymphoid tissues, extracellular calcium sensing stimulates B-cell activation and function through activation of various pathways such as PI3K, NFκB and Toll like receptor pathways [188].

In CLL, a role of *PCLO* mutations in augmented calcium responsiveness has been suggested. Yet, few studies have addressed the functional consequences of mutated *PCLO* gene generally in cancer and specifically in CLL. However, increased calcium responsiveness is found to be associated with shorter lymphocyte doubling times (LDT) and shorter time to treatment [189]. Mutations in *PCLO* have been found in up to 9% of representative CLL cohort included 11 untreated and 34 relapsed/refractory CLL cases [190] and up to 11% of CLL cases with Richter transformation using NGS techniques [191]. Moreover, very common mutations in this gene were observed in up to 35% of patients with denovo DLBCL [192].

1.2.4.3.11. *LRP1B*

This gene located on chromosome 2q21.2. It composed of 91 coding exons. It encodes a member of low density lipoprotein cell surface receptor family of 4599 amino-acids which possesses four ligand binding domains in the N terminal part. Through interaction with multiple ligands, the gene exhibits a diverse biological role in lipoprotein catabolism, cell development, cell adhesion and migration. Furthermore, it is a putative tumour suppressor gene and one of the most commonly mutated genes in a number of human cancers [193]. Mutations in this gene have been found to be associated with disease progression and

metastasis in hepatocellular carcinoma [194], lung cancer [195] and with drug resistance in ovarian cancer [196]. Whole exome study of CLL found mutations in *LRP1B* in 5% of patients [149]. The functional significance of this gene and the consequences of its alteration in CLL are still unclear.

1.2.4.3.12. *HIST1H1E*

This intron-less gene is located on chromosome 6p22.2. It encodes a histone linker protein named H1e, which with other histone linker subtypes namely H1a, H1c and H1d play a fundamental role in the regulation of gene expression at the chromatin level. Conjoining of the histones (fundamental nuclear proteins on which DNA is wrapped around) leads to compaction of chromatin. It is the degree of this compressed chromatin structure that controls the DNA-binding transcription factors to access their cognate binding sites [197].

HIST1H1E mutations occurred in 4% of CLL patients using a WGS approach in 145 unselected CLL patients [198]. The functional significance or consequences of its mutations in CLL is still unclear. However, it is among the putative driver gene mutations identified recently in a comprehensive analysis of WES data from 538 CLL cases of three trials [93].

1.2.4.3.13. *ZFPM2*

This gene is positioned on chromosome 8q23. It has 8 coding exons encoding zinc finger protein (1151 amino-acids) which is a member of the FOG family of transcription factors. It plays a pivotal role in modulating the activity of GATA family proteins (important regulators of haematopoiesis) [199]. Alterations in this gene have been found to affect 5% of CLL patients [190].

1.2.4.3.14. Other novel recurrently mutated genes

Numerous other recurrently mutated genes have been identified in CLL. The frequency of these mutations is generally below 4% for each. Some of the most important genes that are involved in pathways implicated in CLL pathogenesis are shown in Table 1.1.

1.2.5. Clonal evolution in CLL

1.2.5.1. ATM-P53 pathway in clonal evolution

The ATM-p53 pathway constitutes a meshwork of genes that are targeted to respond to a variety of intrinsic and extrinsic cellular stress signals that affect homeostatic mechanisms to monitor DNA replication, chromosome segregation and cell division [319]. Briefly, in response to a stress signal such as irradiation subsequent DNA double strand breaks emerge. This damage brings about phosphorylation of ATM which then activates p53 protein in a specific manner by post-translational modifications, including phosphorylation. The activation of p53 leads to initiation of DNA damage repair, arrest in cell cycle, a program that induces cell senescence and/or cellular apoptosis [320]. In cancer cells with defective ATM-p53 pathway, various intrinsic or extrinsic stresses may result in abnormal DNA replication, genome instability and cell cycle progression which favour survival and proliferation of these cells and consequently result in clonal evolution of the cancer and deterioration of clinical status of the affected individuals. Figure 1.1. shows a summary of ATM-p53 pathway in response to DNA damage.

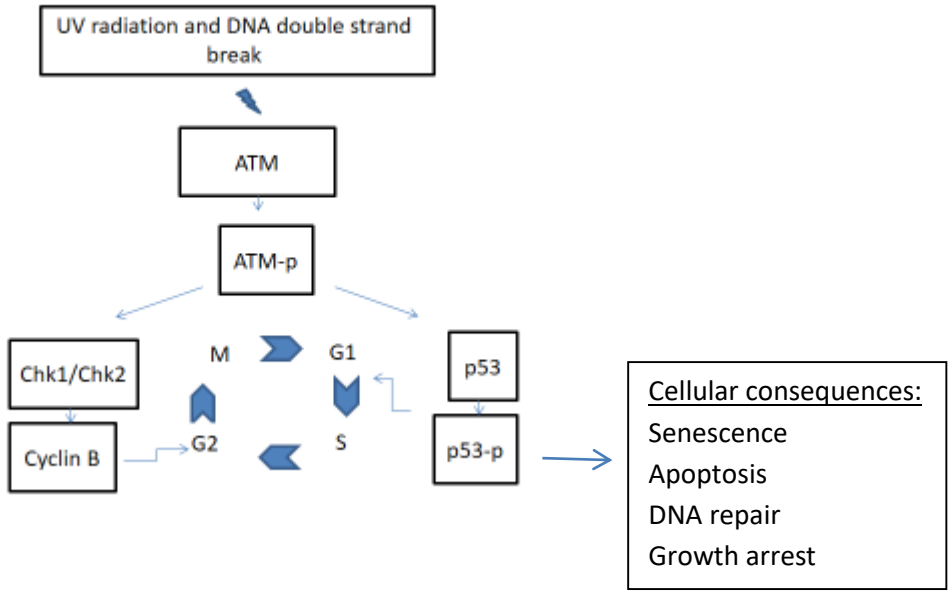


Figure 1.1. A summary of **ATM-P53 pathway.** –p: phosphorylated proteins; G1>S>G2>M: 4 phases of cell cycle; Chk1: Checkpoint kinase 1; Chk2: Checkpoint kinase 2;

1.2.5.2. Concepts: models of clonal evolutions based on mutation profiles

Clonal evolution is one of the enabling characteristics of CLL, in which the acquisition of successive multiple mutations result in alterations in its genetic composition (make up) overtime. This depends on positive selection of additional somatic mutation(s) that are able to compete and adapt to external pressure such as chemotherapy. Therefore, the phenotypic change of CLL (disease progression and relapse) must stem from underlying genetic evolution. This concept has indeed been confirmed in the past by FISH or other cytogenetic techniques; For example, clonal evolution was identified in a considerable number of patients (up to 43%) using FISH or other cytogenetic techniques (of low sensitivity), with frequent acquisition of the poor prognostic markers such as del (11q) and del (17p) over time of progression and relapse [200].

More recently, while the current study was ongoing, models of CLL clonal evolution have been proposed through application of high throughput sequencing techniques (mainly WES) for longitudinal samples. Firstly, linear evolution is characterised by increase in mutant clone population with preservation of the initial genetic composition with or without maintenance of clonal equilibrium. Secondly, branching evolution in which the genetic composition is altered in the progeny such that the new developing clones carry mutations diverse from the ancestor clones [198]. Based on mutation profile, the founder of a clonal mutation (exist in ≥ 95% of the tumour cells) should have occurred before or during the latest selective step, while subclonal ones should have occurred after the most recent selective step [198].

Table 1.1. **Summary of some other recurrently mutated genes with low incidences**

Pathways involved	**Gene symbol**	**Chr. location** (Ensembl)	**% of cases**	**CLL cohort type and No. of patients**	**Method**	**Ref.**
MAPK-ERK pathway/BCR signalling	*MAP2K1*	15q22.31	2%	Unselected (n=287)	WGS	[93]
	BRAF	7q34	3.7%	Unselected (n=160)	WES	[198]
	NRAS	1p13.2	2%			
	KRAS	12p12.1	2.6%			
	PLEKHG5	1p36.31	3%	Unselected (n=48)	WES/n=5 Targeted NGS/n=48	[155]
RNA processing	*DDX3X*	Xp11.4	2-3%	Unselected (n=81) Unselected (n=160)	WGS + WES WES	[150, 198]
Toll like receptor/inflammatory	*IRF4*	6p25.3	1.5%	Untreated and relapsed (n=45)	Targeted NGS	[190]
	MAPK1	2q11.21	3%			
Wnt/MYC signalling	*MED12*	Xq13.1	2-5%	Unselected (n=160)	WES	[198]
	FUBP1	1p31.1	1.7%			
	MGA	15q15.1	3.2%	Unselected (n=287)	WGS	[93]
Chromatin modification	*ASXL1*	20q11.21	2%	Unselected (n=287)	WGS	[190]
	ZMYM3	Xq13.1	2%			
	IKZF3	17q21.1	2%	Unselected (n=287)	WGS	[93]

1.2.5.3. Existence of subclones prior to CLL progression or drug resistance and the importance of their identification as early as possible

Subclonal mutations are expected to occur at early stages of CLL. These subclonal mutations more commonly follow a complex branched pattern of evolution rather than a traditional linear model of evolution. This is proposed to be due to complex interaction among highly diverse population which by itself fuels clonal evolution. Importantly, it is the evolution of mutant subclones that is responsible for disease development in CLL. Therefore, diverse subclonal composition can predict poor prognostic outcome such as disease progression and treatment unresponsiveness [198]. As found in recent studies, subclonal mutations in *TP53* gene have similar effects of clonal mutations in terms of disease progression and response to treatment [138]. Recurrent subclonal mutations in other genes including *ATM, NOTCH1* and *SF3B1* have been also found to be associated with disease progression and/or chemoresistance in CLL [137, 201].

1.2.5.4. A need of sensitive methods to detect small subclones with driver mutations

It is important to note that to date majority of the published data in CLL have been limited to the detection of gross clonal heterogeneity (clone size > 10% of the entire cell population) [198]. This is because only clones that existed in a substantial proportion of the CLL cells or clones that become dominant at the time of study were traceable using the less sensitive methodologies [137]. So that, achieving deeper sequencing depth is important to enable identification of smaller subclones and thus maximising the ability to decipher the true extent of genomic heterogeneity and to provide early prediction for disease progression and/or resistance to therapies in this disease.

1.3. Methods for identification of genomic aberrations in CLL

1.3.1. Identification of chromosomal abnormalities and copy number changes

In the early 1950s, chromosome banding following the discovery of hypotonic solution heralded the recognition of exact chromosomal number in human cells. Abnormal numbers or structures of chromosomes were noticed in various human diseases including cancers with high mitotic activity [202]. In the 1970s, further refinement of this method followed upon the discovery of mitogens. This discovery was particularly useful to study the CLL genome which is characterised by low mitotic activity. Although chromosomal abnormalities could be detected in this disease, the incidence and complexity of these aberrations was underestimated in this tumour [202].

Subsequently, in-situ hybridisation was developed in which specific emission probes of thymidine were used to detect DNA fragments. However, this was a complicated and time-consuming test which required experienced hands and storage of the fixed slides in light proof boxes for 2 months to obtain results. It was therefore not an ideal test for clinical diagnosis. Eventually, a successive development of florescent probes for FISH has circumvented this problem [202].

1.3.1.1. FISH

FISH was developed in 1980s. It uses florescent probes to detect complementary DNA or RNA sequences, and is therefore more specific, sensitive and time-saving compared to other conventional cytogenetic techniques. With its ability to identify chromosome rearrangements in non-dividing cells and further refinement for using multi-colour probe painting, it has become a particularly useful tool to identify prognostic subgroups in haematological malignancies [203]. In CLL, interphase FISH has been successfully used to detect recurrent chromosomal aberrations (e.g. del 17p, del 11q, del 13q and trisomy 12) which remarkably help in identifying high-risk disease. However, chromosomal aberrations

detected by FISH are incomplete as limited chromosome regions being targeted and often difficult to be quantified.

1.3.1.2. Array based karyotyping methods

Comparative genomic hybridisation was the earliest array method used for karyotyping studies. The patient DNA and a normal control DNA are differentially labelled with florescent dyes which are then subjected to competitive hybridisation with a normal chromosomal preparation. Colour deviation of the test sample from the control is then used to computationally analyse the results.

Although initially the method produced a picture of chromosomal abnormalities across the genome without the need for metaphase spread, only large gain or losses were detectable owing to its low resolution. Application of oligonucleotide or bacterial artificial chromosome potentiated the detection of smaller deletion or amplification spanning in Kilobases. It is also possible to study the entire genome or to target specific locations within the genome. Its main drawback is inability to detect balanced translocations [204] .

SNP array is the latest developed method for detecting genome-wide CNA. Various SNP array methods with different resolution have been developed depending on the number of probes specific to common single nucleotide polymorphisms spanning a particular chromosomal region. This technique allows detection of copy neutral loss of heterozygocity (LOH) or uniparental disomy (CNN-LOH) that is loss of one homologue of a chromosomal region and replacement of this loss by doubling of identical segment of the other homologue. An example of high resolution SNP array is Infinium CytoSNP-850K BeadChip which utilises single-base extension. Bead bound DNA probes are attached to the array surface (each bead is attached by identical copies of the same probe). Amplified fragments of genomic DNA are hybridised to the probes. Biotin-labelled ddCTP and ddGTP, and 2, 4-dinitrophenol (DNP)-labelled ddATP and ddUTP are incorporated into the probe sequence depending upon the genotype at that base. The hybridised DNA is then detached and washed out to leave the fluorescent labelled probes. Probes then receive a fluorescent label;

streptavidin labels biotin green and anti-DNP antibody labels DNP red. After completion of the assay, the chip is scanned by a laser and the intensity of the fluorescence is measured. The relative intensity of red and green fluorescence is dependent upon the genotype [204, 205].

In CLL, copy number changes are commonly used as prognostic indicators. For example, for patients with 17p deletion, a different regimen or a more intensive treatment might be selected. Conversely, for patients with isolated 13q14 deletion or normal karyotype, "watch and wait" approach might be used in management plan. As mentioned earlier, array-based karyotyping is superior to FISH and CGH in performance. This is because it has higher resolution and greater coverage of genomic regions. It has been found that 50% of patients who had been assigned as normal karyotype for regions covered by FISH probes, had subclonal population of genomic lesion with prognostic significance outside the detection limit of FISH. Furthermore, SNP array methods detect both copy number abnormalities and copy neutral events, such as uni-parental disomy (UPD) and genomic complexity that cannot be identified by FISH and CGH. Moreover, acquisition of additional genetic lesions is a sign of clonal evolution which has been shown to shorten overall survival of a median of only 22 months [206].

UPD has been reported in CLL patients with *TP53* mutations, thus 2 mutated copies have been produced which has the same clinical significance of 17p del in CLL risk assessment. Moreover, the q arm of chromosome 13 has been reported to be affected by UPD which changed the original heterozygous deletion at 13q14 to a homozygous deletion with identical break points on both alleles [206, 207]. More recently, the utility of whole genome and whole exome NGS data has been extended to detect copy number changes, thereby they can serve as substitutes for identification of SNVs and CNAs [208].

1.3.2. Identification of point mutations, small insertions and deletions

There are various methods for detection of point mutations and small indels. The choice of a method depends on the status of mutation (whether it is known or unknown), number of the mutations, and size of DNA regions where the mutations exist and reliability of the method.

1.3.2.1. Single strand conformational polymorphism (SSCP)

Various tests can be used to detect unknown point mutations using the physical property of denatured DNA. For example, SSCP uses the concept that DNA variations cause alterations in the conformation of denatured DNA fragments. It compares the altered migration of denatured wild-type and mutant DNA fragments during the electrophoresis [209]. In this technique, DNA fragments up to 200 bp can be screened. It has the capacity to detect up to 80% of potential point mutations or any sequence change (synonymous or non-synonymous single nucleotide variation as well as short indels) [210].

1.3.2.2. Heteroduplex formation

Likewise, heteroduplex formation by mutant DNA and its differential mobility as compared to homoduplex forming wild type DNA is the basic concept of using denaturing high performance liquid chromatography (DHPLC) to detect point mutations [211]. Similarly, based on differences in the melting behaviour of DNA even with single base substitution, denaturing gradient gel electrophoresis (DGGE) has been used. This technique is beneficial because DNA fragments up to 1Kbp can be screened; moreover, over 95% of point mutations can be reliably detected by this method. Although this method is considered to be labour intensive [211].

1.3.2.3. Property of mutant proteins

For some genes, the mutated proteins can be used for unknown mutation screening if their properties are predictable. The well-known example is to use mutant p53 to identify samples harbouring *TP53* mutations in human cancers including CLL.

1.3.2.3.1. Mutant p53 protein expression

In contrast to wild type p53 protein which stays too short to be readily detected in unstimulated normal cells, the mutant p53 has a prolonged half-life [212]. Therefore, the high baseline level and lack of activation of p53 protein as measured with either immunohistochemistry or flow cytometry suggest *TP53* mutation in CLL [213]. Nevertheless, its clinical utility is compromised due to false negative and false positive from stop codons

and destabilising mutations. Additionally, result variability from technical aspects such as the use of different antibodies and scoring criteria were frequently encountered [212].

1.3.2.3.2. Loss of transcriptional activity

Due to the disadvantages of the above method, a technique was established to detect any unknown mutations affecting the transcription activity of p53 which is called functional analysis of separated allele in yeast (FASAY). The basic principal of this assay is that p53 functions in gap repair by homologous recombination and its transcription activity can be detected in the yeast report system. In this test, TP53 cDNA (exon 4 - exon 9) is inserted into a linearised vector followed by co-transformation in to yeast cells. Identification of functional p53 protein depends on the yeasts' minimal promoter gene expression upon which white colonies are produced. While red colonies are produced if mutated p53 are inserted because of lack of the promoter gene expression [214].

This method has been used as an initial screening test for CLL cases to identify *TP53* mutations followed by Sanger sequencing of the colonies to determine the location and type of nucleotide change because of its cost effectiveness. The main advantages of this method are; firstly, the ability to analyse both alleles separately; secondly, a substantial part of the gene can be quantitatively analysed with a high sensitivity; thirdly, fully inactivating mutations can be distinguished from partially inactivating ones; and lastly, the colonies can be sequenced for confirmation [214]. On the other hand, the method is technically demanding and time consuming. Moreover, it cannot detect mutations causing alternative splicing or mutations that are transcription independent. Besides, only the coding sequences between codons 42- 374 of the gene are examined [214].

All the above methods described are designed for screening unknown mutations. The following methods are used to detect known mutations.

1.3.2.4. Allele-specific PCR

A type of polymerase chain reaction (PCR) called allele specific (AS) PCR can be used to detect known point mutations or small indels. In this method, two PCR reactions are performed using respective primers specific for mutant and wild-type sequences.

The method is valuable because in addition to identifying the genomic state (wild type or mutant), the carrier status (heterozygous or homozygous) of a sample can also be determined. The method is sensitive to detect mutation with allelic frequency as low as 1% [215]. However, its utility is hampered as it is restricted to known mutations and there is a frequent occurrence of non-specific amplifications, which require use of known control samples and optimisations [216].

1.3.2.5. Restriction fragment length polymorphism (RFLP)

RFLP uses the concept that some point mutations can change restriction sites in DNA causing alteration in cleavage by restriction endonuclease enzymes which produce fragments with various sizes. Therefore its use is limited to detecting mutations that occur in restriction sites [210].

1.3.2.6. Sanger sequencing

Sanger sequencing is the first generation DNA sequencing technique. It was developed by Frederick Sanger and colleagues in 1977 and is still a gold standard of sequencing technology as served in the 1000 Human Genomes Project [217].

This technique depends on chain termination. Radioactively or fluorescently labelled single nucleotides are added by DNA polymerase. Although the radioactive labelling and autoradiography were initially used for visualising DNA sequence, the fluorescent dye-terminator sequencing is now the mainstay in automated sequencing owing to its greater expediency and speed. In this modified method, the emitted lights are excited by laser in the DNA sequencer while passing through the detection region. The DNA sequence is revealed from the order of the fluorescent fragments [218]. Subsequently, base-calling and error probability assignments applications are used to call the DNA sequence and evaluate

the accuracy of the base calling which is scaled by Phred score. Any base that achieves the standard Phred 20 (Q20) is a high quality base as its error probability is equivalent to 1% [218]. Through this and other further improvements, sequencing reads of up to 900 bp can be now obtained with high (99.999%) accuracy [219]. Conversely, because the sequencing is performed on four individual amplicons, each of which contains only one of the four dideoxynucleotides (ddATP, ddCTP, ddGTP or ddTTP), variation in intensity of fluorescent emission can result in less precise signal recognition. This appears as misleading small background peaks. Therefore its sensitivity is limited to identification of mutations with allele frequency greater than 10% [218].

Sanger sequencing bears technical limitation for sequencing large number of targets as it is not time or cost efficient. It requires PCR amplification for DNA template (library) preparation which usually involves either DNA cloning or gel purification. The estimated cost for an 800 bp sequencing reaction is £3. Another big challenge for this method is that it can directly sequence only relatively short (300 - 1000 nucleotides) DNA fragments in a single reaction. All of the above limitations hamper its further application and justify development of new techniques with a higher capacity and efficiency to sequence a big size of DNA region in a large number of samples concurrently [217].

1.3.2.7. Next generation DNA sequencing

Next generation sequencing (NGS) is a new technique developed only a decade ago. It is characterised by massively parallel sequencing in which up to millions of DNA fragments from a single sample or multiple samples are sequenced simultaneously. Unlike Sanger sequencing, this technology has witnessed improvement in mutation detection at higher sensitivity based on the large order of sequencing magnitudes each target nucleic acid can achieve through deep sequencing approaches [220]. Various NGS platforms are commercially available and newly emerging platforms are continuing to be developed. These platforms have different throughput capacity. Large scale machines such as Illumina Hiseq can generate Gigabases (Gbp) of sequences per run which allow an entire genome to be sequenced in less than 4 days. More recently, Illumina HiSeq X10 is underway which produces 1.8 Terabases (Tbp) of sequence [221]. This tremendous increase in throughput

allows the whole human genome to be sequenced at reduced cost; however the system itself is very expensive which limits its availability only to large institutes performing population-scale genome sequencing [221].

Lower throughput bench top sequencing machines such Roche 454, Illumina Miseq and Ion Torrent (more recent) are also available. They can generate Megabases (Mbp) of sequencing output at relatively low cost. This has made sequencing facilities accessible to more labs and highlighted an interest for their use in a clinical setting which require cheaper, faster, and easier-to-use sequencer [222]. Various platforms capacity and sequencing mechanisms are summarised in (Table 1.2).

Table 1.2. **Comparison of the most widely used NGS platforms** [223, 224]

Platforms parameters	**Roche 454 GS**	**SOLiD 5500 xl**	**Illumina HiSeq 2000**	**Illumina MiSeq v1**	**Ion Torrent PGM**	**PacBio RS**
Amplification	Emulsion PCR	Emulsion PCR	Bridge amplification	Bridge amplification	Emulsion PCR	No PCR
Sequencing mechanism	Pyrosequencing	Oligonucleotide ligation	Sequencing by synthesis	Sequencing by synthesis	Sequencing by synthesis	Single molecule real time sequencing
Detection method	Light detection	Light detection	Light detection	Light detection	H^+ ion detection	Light detection
Output/run	50 Mbp	150 Gbp	600 Gbp	2 Gbp	20 Mb-1 Gbp	100 Mbp
Turnaround time	10 hr	8 days	3 days	26 hr	2-5 hr	2 hr
Av. Read length	400 bp	75 bp	100 bp	150bp	100-200 bp	900 bp
Approximate sequencing cost/Mbp	£6	£0.05	£0.03	£0.05	£0.06	£0.12

Although different platforms use different techniques, they commonly share a similar order of workflow (Figure 1.2), including DNA library and sequencing template preparation, data production and data analysis. Among these stages, data processing and analysis are generally most difficult owing to the vast amount of data generated, limitation of informatics infrastructures and time requirement to produce comprehensive analysis [200].

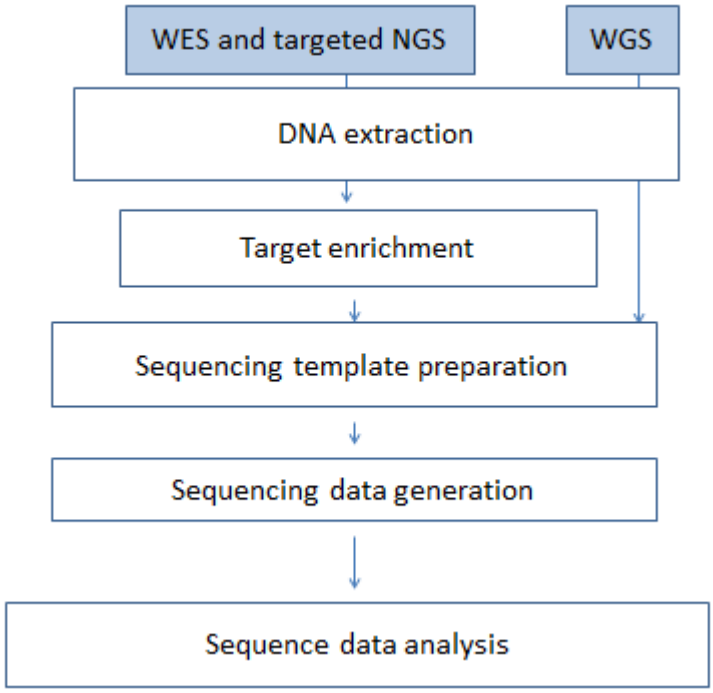


Figure 1.2. **Main stages of next generation sequencing**
The targeted NGS consist of 5 major steps, named genomic DNA extraction, target DNA enrichment, sequencing template preparation, sequencing data production and data analysis.

1.4. Next generation sequencing approaches

The use of next generation sequencing (NGS) has expanded the knowledge of the genomic alterations in CLL as it has led to discovery of new recurrently mutated genes which were previously unknown in CLL. Functional and clinical studies of these novel gene mutations have discovered new mechanisms implicated in the pathogenesis of the disease, revealed new insights into CLL molecular evolution that could ultimately translate into improvements

in the clinical management of patients. There are three NGS approaches, each suiting a specific purpose. A brief description of each approach is outlined in the following sections.

1.4.1. Whole genome sequencing (WGS)

In this approach, the whole human genome is sequenced. Human genome is composed of approximately 3.2 Gbp of nucleotides. It has been feasible to study the genome at population scale grounded by the advancements of NGS in throughput and reduction in cost. The most widely used platforms are Illumina HiSeq 2000 (100 Gbp), Illumina HiSeq 2500 (1000 Gbp) and the most recent platform Illmina HiSeq X Ten (10000 Gbp) marketed in 2014 [225, 226]. The average coverage sequence depth is usually 30 - 50 x.

Since the accomplishment of the first large-scale human genetic variation study namely the 1000 Genomes Project in 2004, larger unprecedented projects have been launched. These projects include the sequencing of thousands of genomes which are disease specific. Whole genome sequences enable understanding of the relationship between genomic variation and phenotype. Moreover, it serves as comprehensive testing in clinical diagnosis of various types of genetic disorders including identification of novel genetic factors particularly in inherited diseases [227]. However, this approach is still unpractical because of many facts; the cost remains relatively high, the turnaround time from the sequencing itself and data analysis is still pronounced and data storage is a challenging issue [221]. Moreover, its limited coverage depth (range 30 x - 100 x) and low sensitivity render the method inappropriate for identification of somatic alterations carried by small population of cells.

1.4.2. Whole exome sequencing (WES)

This approach has emerged to study defined regions of human genome in which only the coding regions of the genome are sequenced. The protein coding regions comprise nearly 1% of the entire human genome or ~30 Mbp which split across ~180,000 exons [228]. However, this part of human genome contains approximately 85% of known disease- related variants [221]. For this reason, sequencing of whole exome has been more extensively used compared to whole genome sequencing for clinical studies in the recent years [221].

In this approach, only the coding regions are sequenced and therefore the genomic alterations detected reflect sequence changes that result in functional alteration of proteins. Thus the smaller genomic regions targeted result in higher coverage depth usually of 100 x-160 x and therefore increasing sensitivity of identification of somatic changes. In CLL, this approach has been implemented by various groups to study CLL genomic landscape as well as monitoring recurrent gene mutant clonal evolution in different selected and unselected cohorts [149, 150, 229].

1.4.3. Targeted DNA sequencing

In addition of the cost inefficiency and time apprehension in data production and management, the two earlier approaches suffer from lower coverage depth. Therefore, some important disease related variants may be missed through the bio-informatics algorithms applied to call high quality variants. Higher quality variants are achieved through increasing the read depth by focusing on disease specific variants or genes. For many diseases, there are a limited number of variants or genes which can be targeted with better coverage [230]. Not surprisingly, targeted NGS has recently been applied in more and more clinical diagnostics for various tumours including CLL [231].

The key step in targeted NGS is to enrich DNA sequence of interest through library preparation. Target capturing is accomplished by fusing randomly fragmented pieces of DNA molecule to platform specific adaptors. This is followed by PCR to amplify the adapter bound DNA. Later, by the help of the adaptors' immobilisation site and sequencing primer site, the DNA libraries are amplified and sequenced. The adaptors also contain barcodes which are used for indexing multiple samples. Several systems for target enrichment have been developed. Their performances are measured by a number of parameters including sensitivity, specificity, reproducibility and uniformity of coverage. Moreover, cost, ease of use and the amount of DNA consumed are the other important factors specifically for handling clinical samples [232].

Sensitivity of enrichment is equivalent to the percentage of the targeted bases represented by at least a single sequencing read, while specificity denotes the percentage of sequence

reads that map to the target regions. Uniformity is the degree of evenness in sequence coverage across the target regions. Reproducibility is consistency of obtained results from replicated experiments [232]. Here we describe the most widely used target enrichment approaches and summarise the performance of each system.

1.4.3.1. Target enrichment systems for next generation sequencing

PCR based systems depend on primer pairs to amplify the target regions in the genome. To increase the throughput particularly for large target regions, a large number of amplicons must be generated. This is difficult to be achieved by uniplex PCR. Despite multiplexing of PCR through concurrent use of numerous primer pairs for amplification, the occurrence of multiple amplifications often cause interactions between primers which result in a limited number of amplicons successfully work simultaneously. To overcome those difficulties, other methods have been developed for target enrichment.

Emulsion PCR has proven to be useful to overcome the nonspecific amplifications as each emulsion droplet which is occupied by a single template and the primers. However, emulsion PCR fails to produce long amplicons due to the restricted size of the droplets [233]. To overcome this problem, overlapping PCR for adjacent regions has been suggested. However, it is time consuming as individual PCR reactions need to be as efficient as possible to reduce the total amount of consumed DNA. Moreover, after recovery of the PCR products, normalisation is required to circumvent sequencing one dominant product over the others [232]. Altogether, PCR based methods are expensive, difficult to use and require high amounts of DNA which is estimated to be 8 µg for 1 Mbp of target region. Moreover, it loses specificity when targets of large sizes are amplified [232].

Molecular inversion probe (MIP) system works through circularising target DNA regions. The probes are single stranded oligonucleotide sequences. The central portion of the probe is composed of a common sequence so called linker while its two ends are complementary to specific sequences of target DNA. Upon annealing and ligation of the probes with the target, circularisation of the region of interest occurs. The reaction then subjected to enzymatic

digestion to break down un-circularised DNA. The circular DNA segments are then amplified by PCR.

Compared to PCR based systems, this system is less expensive if large numbers of samples are processed. It is not cumbersome, as little as 200 ng of DNA is required and it offers more specificity with large target regions [232, 234].

Hybrid capture systems are available in 2 forms either through array surface capture or in solution capture. In the array system, the target specific oligonucleotide probes are fixed on an array plate. After hybridisation of targeted DNA, the non-specific DNA hybrids are washed out. Various arrays with different capacities have been developed. Unlike PCR based methods, this approach is quicker and less laborious; however it is more expensive and consumes higher amounts of DNA irrespective of the size of the target region [232].

In solution hybrid capture uses DNA probes designed to target region of interest. In addition to the advantage of sample multiplexing and a greater capacity to capture large target regions, it is found to be superior to all the above mentioned systems. This is because it requires less DNA as it uses excess probes over DNA. Moreover, it is easy to use and yields more specific and uniform coverage than array based methods [232]. Examples are the SureSelect hybrid capture method and HaloPlex target enrichment systems. The latter is the newest version of in solution hybrid capture developed by Agilent Technologies in 2012 [235]. It is the enrichment system that has been used in this study. It is more advantageous than SureSelect in that it requires less starting DNA; it is highly specific and provides more even genome coverage. On the other hand, a maximum of 5 Mbp of custom designed regions can be enriched, which is much less compared to 24 Mbp for SureSelect hybrid capture [234]. However, this limitation does not affect this project since the total size of DNA targets is much smaller than the maximum. Therefore, HaloPlex was selected for the current study shortly after it became available.

1.4.3.2. Next generation sequencing data analysis

Analysis of huge amount of sequence data is still the bottle neck in NGS studies. This requires support by bioinformatics, which generally includes a data analysis pipeline, a database for annotation and reporting processes. NGS data analysis is usually composed of a three step computational framework, namely primary, secondary and tertiary stages.

In the primary stage, sequencing signals are processed to produce raw sequence data. This is followed by identification of variants by mapping of the sequence reads to a reference genome. Finally, annotation and filtering of variants is executed. At each stage, an opportunity for quality control is provided to minimise the chance of false variant calls. In-house or platform specific pipelines and external tools are required to process the data [236]. The output data from the primary stage of analysis is in a file of sequence information and quality score (FASTQ). The presence of the quality score helps trimming or filtering of poor quality reads.

In the secondary stage, different mapping algorithms are integrated with flexibility in order to identify the best candidate mapping region and position of sequence variations. The output data file from this process is a Sequence Alignment Map (SAM) which can be compressed to a Binary Alignment Map (BAM) file using specific tools. These file formats can be uploaded to a freely available source called Integrative Genomic Viewer (IGV) to visualize the coverage and sequence of the aligned reads [237]. The output of the variant calling process is used to generate Variant Call Format (VCF) file that includes information on the variant's chromosomal location and its quality metrics.

In the last stage, VCF files are exported to external software for identification of common germline polymorphisms. This identification is possible because all the external software have integrated genomic mutation databases such as dbSNP and 1000 Genomes Project. Some software also have incorporated prediction tools to identify the potential significance of each variant [236]. The comparison to known databases helps to classify variants identified by NGS into different types, e.g. germline or somatic, biologically tolerated or deleterious, and clinically benign or damaging alterations.

1.5. Outline of targeted NGS using HaloPlex and Ion Torrent PGM

Since HaloPlex is used to enrich target DNA and Ion Torrent PGM for the sequencing in this study, the details of these techniques are introduced in the following sections.

1.5.1. Steps in target enrichment with HaloPlex

There are five major steps in this system as presented in Figure 1.3. Firstly, genomic DNA is digested by sixteen restriction enzymes to form DNA fragments with a size ranging from 150 - 550 bp. Secondly, the fragmented DNA molecules are circularised and hybridised to biotinylated probes which at both ends, are complementary to the target DNA. Thirdly, through ligation and multiple steps of washing the on-target DNA is retrieved by streptavidin coated magnetic beads while the off-target DNA fragments are removed. Fourthly, PCR is performed to amplify the captured DNA using universal primer complimentary to a sequence in the probes' linker region. Lastly, size selection of the PCR product is performed using solid-phase immobilisation AMPure XP beads which is carboxyl-coated magnetic particles. These particles can reversibly bind to negatively charged DNA in the presence of polyethylene glycol (PEG) and sodium chloride. This allows the free adapter, PCR primer dimers, as well as short PCR products (e.g. < 100 bp) to be removed from the library [238].

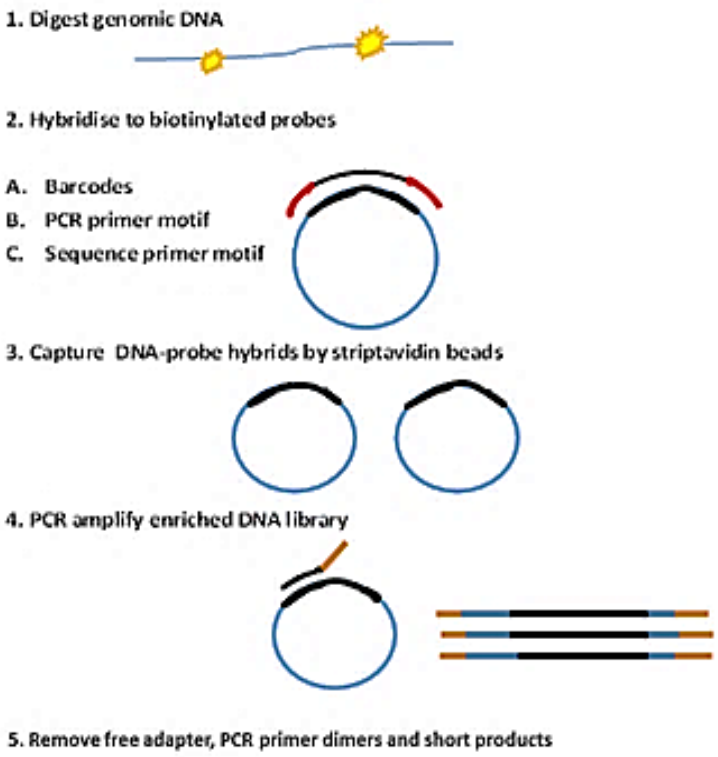


Figure 1.3. **Target DNA enrichment by HaloPlex system**
This system utelizes 8 sets of restriction digestion enzymes to fragment the genomic DNA. The fragments are then circularised by hybridisation to biotinylated probes which contain sequences complimentary to target DNA at both ends, barcodes, PCR primers and sequence primers. The target DNA is captured by streptavidin beads and the ends are then ligated. Only the properly ligated DNA is kept after multiple washing steps, which is subjected to PCR amplification. Following the PCR, size selection is performed to clean up the amplicons from free adaptors and primers.

Following these major steps, quality of the enriched DNA library is verified using Bioanalyser or tape station. The quantified DNA libraries are then pooled if barcoded before being sequenced on a specific platform.

1.5.2. Sequencing using Ion Torrent PGM

Ion Torrent PGM became available in the market at the end of 2010, shortly before the start of this study. As introduced in section 1.3.2.7, it is a small bench top machine which uses

semiconductor sequencing technology instead of optical light detection. So far, it is the fastest and cheapest NGS platform.

1.5.2.1. Versions of Ion Chip and their capacities

Three different pH sensitive chips were available at time of this study; version 1 of Ion 314, 316 and 318 Chips were replaced by version 2 for corresponding chips with improved total output. These chips have different output depending on the number of micro-wells on the surface of the chip (Figure 1.4) [239]. The higher the number of micro-wells, the higher the number of reads generated. Although the sequence output is also dependent on the sequence read length, the read length is largely dependent on the sequencing template preparation kit rather than the type of chips [240].

Figure 1.4. **Capacity scales of various Ion Torrent chips available at time of this study**
The number of micro-wells on the surface of the elliptical shaped area is shown for each chip. The sequence output depends on the type of chip used as well as on the type of the sequencing template preparation kit**.**

1.5.2.2. Sequencing template preparation

Initially, unique DNA library molecules are attached to the surface of beads (Ion Sphere Particles (ISP)). The DNA library molecules on the beads are then clonally amplified in emulsion PCR reaction. The emulsion droplet sequesters clonally amplified libraries on the surface of the beads. This process is performed using automated One Touch systems.

1.5.2.3. Chip loading and signal generation

Following the above steps, the ISPs with amplified DNA library are loaded on the Ion Chips. Through centrifugation, one ISP particle is deposited into each micro-well in the chip. Cyclical flooding of the wells with nucleotides automatically occurs. As presented in Figure 1.5, in each cycle, the incorporation of a nucleotide to the DNA molecules by the polymerase results in release of a hydrogen ion (H^+) and reduction of pH [239]. Through detecting the difference in pH by the underlying pH sensor, the machine recognises whether the nucleotide is complemented or not. The pH shifts are then changed to electrical voltages. Therefore, no voltage will be found if no nucleotide is added while if 2 nucleotides are added, double voltage peaks are detected. However, adherence to this rule is diminished in the presence of long homo-polymeric sequences [219].

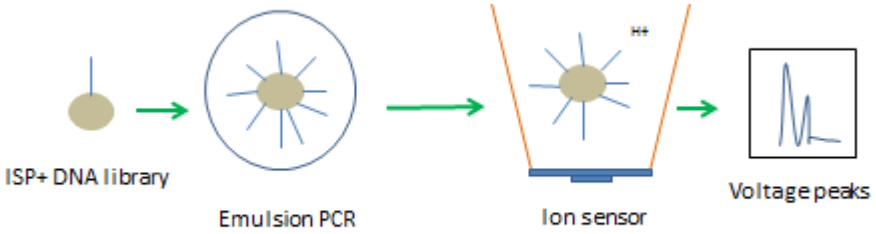


Figure 1.5. **Sequencing template preparation by emulsion PCR for Ion Torrent PGM** In optimal emulsion PCR conditions, each ISP is attached by one DNA library fragment which then encircled by an emulsion droplet to clonally amplify the library on the surface of ISP. The ISP then loaded into the micro-well which is underlined by a pH sensitive layer. Upon nucleotide incorporation with cyclic flooding of the micro-wells, an H^+ ion is released which can be detected and converted to electrical voltage and sequencing signals by the machine.

1.5.3. Ion Torrent sequence data analysis

The most commonly used pipeline is Torrent Suite Pipeline which has been developed and optimised by the manufacturer for Torrent sequence data analysis (Figure 1.6). At the start of this project version (v3.2) was available. Later, a newer version (v4.2) was released which differed from the earlier version in that more stringent parameters were introduced to more accurately process signals, align reads and improve variant calling accuracy. The raw signal on the Ion Torrent Server is algorithmically converted in to linear sequence data in FASTQ format by Torrent Base Caller. This file can be exported to various external pipelines for downstream analysis.

Alignment of the sequence reads to a reference genome yields BAM file format by Torrent Mapping Alignment Programme (TMAP). This programme uses a combination of Burrow Wheeler Aligners (BWA) and Sequence Search and Alignment by Hashing Algorithm (SSAHA) to refine the created mapping locations. The run summary contains chip loading, total output, sequence quality, mapping quality metrics and read length is displayed in the Torrent Browser Report summery page. By using the Torrent Variant Calling (TVC) plugin, custom specific algorithms can be applied to precondition variant recognition and calling. The output file generated after this process is Variant Call Format (VCF). This file contains the list of the called variants, their chromosomal location and quality metrics.

The VCF file can be exported to various open access variant effect predictor software or to Ion Reporter™ Software either manually or automatically using Ion Reporter™ Uploader plugin. The automatic uploading operates from the initial step of BAM file processing.

The advantages of using Ion Reporter™ Software are; firstly, time effectiveness as it needs less bioinformatics work to figure out the effects of the identified variants; secondly, the software integrated comprehensive public annotations which enable germline variant filtration; thirdly, it provides an organised workflow for comparing multiple samples.

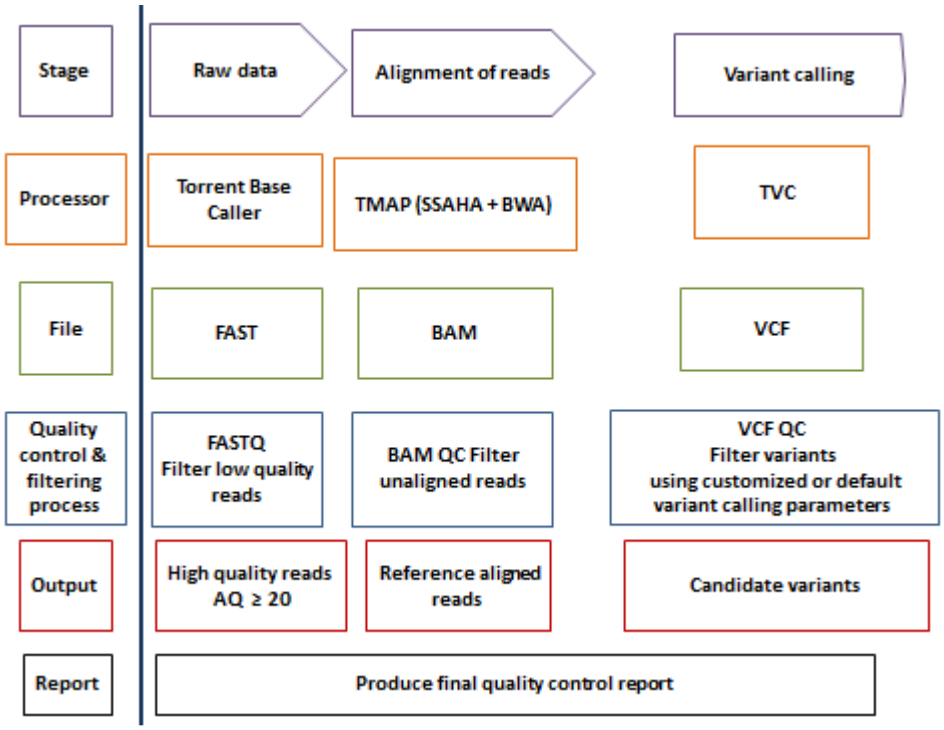


Figure 1.6. **Torrent Suite pipeline work flow**
The processors, quality control and output files in each stage of data processing are shown. The process starts from raw sequence data generation from conversion of electrical voltage signals and ends with variant calling which produce VCF file.

1.6. Sources of error in NGS data

NGS data like any other sequencing technology might harbour false positive or false negative variants. The sources for false positive calls might be PCR duplicates which tend to occur in the same position of unified PCR amplicons or at around the indels. The former is usually filtered out by filtering PCR duplicates, while the latter is usually prevented by specific [322] tools that locally realign the reads around the indels to the reference sequence for example genome analysis tool kit (GATK). More importantly, having accurate and proper reference sequence is important to correctly identify mismatched bases, in addition to the need of germline sequence information to distinguish somatic mutations. On the other hand, false negative results might be encountered if shallow coverage or no coverage of target genes were achieved. Details for avoiding false negative calls will be shown in subsequent chapters of this thesis.

1.7. The study hypothesis and aims

As mentioned earlier, through application of massively parallel sequencing, novel genes have been found to be recurrently mutated in CLL. Some of these genes have been associated with disease course and response to treatment. It has been found that mutations occur at early disease stages and can expand either at the time of disease progression or after chemotherapy through selection of fitter and more resistant clones. However, it is still not fully understood whether and what interrelationships exist among these mutant clones/and or subclones to affect the evolutionary process and contribute to chemo refractoriness.

It is therefore important to identify minor mutant subclones to guide treatment choice and predict prognosis at earlier stages. A sensitive, fast and affordable next generation sequencing technique is needed to meet this clinical demand. This prompted us to conduct this study with aims firstly to develop a deep sequencing method to detect known genomic alterations in targeted multiple genes relevant to biology and clinical outcome of CLL; and secondly to apply it to study evolution of those mutant subclones and their interrelationships in disease progression and resistance to treatment. Ultimately, we hope to develop novel molecular biomarkers for this disease.

Chapter 2. Development of ultra-deep targeted next generation sequencing based on combined HaloPlex target enrichment and Ion Torrent PGM techniques

2.1. Introduction and aim

Shortly before the beginning of this study, a number of novel gene mutations had been identified in heterogeneous cohorts of CLL cases with high throughput next generation sequencing (NGS), including whole genome (WGS) and whole exome (WES) sequencing techniques [241]. Their frequencies in CLL had been revealed and an association with advanced disease and drug resistance had been also implied. In subsequent studies, clonal evolution of chromosomal aberrations and its association with disease progression and outcome were described [117]. Evidence suggested that diverse CLL subclonal populations arising from a common ancestor can exist before and be selected by therapies. Thus highly fit or resistant subclones survive and eventually dominate [242]. These findings highlighted the need for a more sensitive targeted sequencing technique to identify the low level of these identified mutations in order to monitor clonal evolution and to predict disease outcome as early as possible. However, no such technique was available before this study. We were therefore prompted to firstly establish an in-house deep next generation sequencing approach.

For this purpose, HaloPlex was selected for enrichment of target DNA and Ion Torrent PGM as the sequencing platform. HaloPlex target enrichment system was an in-solution hybridisation target capture method launched by Agilent shortly before this study. Given that it requires the least amount of g. DNA, provides scalable target size capture and enables sample multiplexing, it is an ideal candidate method for intended clinical application in the future. Ion Torrent PGM was among the most recently commercialised bench top next generation sequencing platforms. Owing to its scalability in throughput as well as its time and cost effectiveness as described in Chapter 1 (Section 1.3.2.7), it would be the best

nominee to be used for clinical service. Besides, this platform was in-house accessible and provided a software package for automatic data analysis.

The aim of work presented in this chapter, was to step by step optimise conditions and validate results of a NGS deep sequencing method to test for mutations in a CLL gene panel designed for clinical diagnostic service in the future.

2.2. Materials and methods

2.2.1. CLL sample selection, processing and storage

All CLL cell samples used for this study were cryopreserved in the Liverpool Leukaemia Bio-bank with informed consent of patients and the approval of the Liverpool Research Ethics Committee. Diagnosis for all patients met the criteria recommended by 2008 IWCLL [21]. For the purpose of this chapter, we used 12 CLL samples to develop the highly sensitive NGS method. Five of these samples had previous archived DNA sequence information for *TP53*. These included a CLL case with wild-type *TP53* and 4 cases with 4 distinct point mutations in *TP53* as detected by functional analysis of separated alleles in yeast (FASAY) and Sanger sequencing. These mutated samples were used to test the sensitivity of the method by serial dilution of the point mutations with the *TP53* wild type DNA sample. 3 of these *TP53* mutated samples as well as the wild-type DNA sample were also re-screened with the NGS method prior to performing the sensitivity test. Furthermore, 2 CLL cases with wild-type *ATM* and 5 cases with a total of 8 distinct mutations in *ATM* detected by the Ion Torrent PGM were sent to another centre for confirmation. Details of mutations in those samples are shown in Table 2.1.

Table 2.1. **Clinical and genetic information of the CLL samples used in this chapter**

CLL-ID	Gene mutation status	Gene ploidy by FISH	Chr. locations and N. changes (hg19)	VAF %	In-house methods applied
CLL-A	Mu *TP53*	Del 97%	17:7577539G>A	96	FASAY+Sanger
CLL-1	Mu *TP53*	N	17:7576891T>A 17:7577079	24 5	FASAY+Sanger+PGM PGM
CLL-4	Mu *TP53*	N	17:7577100T>C	74	FASAY+Sanger+PGM
CLL-6	Mu *TP53*	N	17:7578394T>C	98	FASAY+Sanger+PGM
CLL-11	Mu *ATM*	N	11:108121763G>A 11:108213973G>A	37 38	PGM
CLL-12	Mu *ATM*	N	11:108186598T>C	98	PGM
CLL-15	Mu *ATM*	Del 64%	11:108143528T>G 11:108168106-7InsC	14.7 14.4	PGM
CLL-17	Mu *ATM*	N	11:108186599A>G 11:108216582-86Del5Nt	47 48	PGM
CLL-18	Mu *ATM*	Del 58%	11:108198445-54Del10Nt		PGM
CLL-22	Wt *ATM*	N			PGM
CLL-24	Wt *TP53*	N			FASAY+Sanger+PGM
CLL-32	Wt *ATM*	N			PGM

Mononuclear cells were prepared from patients' peripheral blood according to a standard operating procedure (SOP) of the Bio-bank. Briefly, the heparinised whole blood was loaded on Lymphoprep ™ (Axis-Shield PoC AS, Norway, d = 1.077) and centrifuged at 800 xg for 30 minutes to collect the mononuclear cells. This was followed by washing and re-suspending the cells in ice cold RPMI-1640 supplemented with 10% fetal calf serum (both from Sigma-Aldrich, UK). These cells were then mixed slowly on ice with the same volume of chilled

RPMI-1640 containing 20% Dimethyl sulphoxide (DMSO) (Sigma-Aldrich, UK) so that the concentration of cells was at 2 x 10^7/ml. Immediately, each 1-ml aliquot of this cell suspension was transferred to a labelled cryo-vial and left at -80 ^oC overnight before being stored in -150 ^oC freezers. With this procedure, purity of CD19+ lymphocytes (approximately equal to CLL cells) were > 90% [243]. Additionally, DNA from 5 CLL cases was received from the University of ULM, Germany to be screened for identification of *TP53* mutations in a multicentre double blinded study organised by the European Research Initiative on CLL (ERIC).

2.2.2. Genomic DNA extraction from CLL cells

After thawing from -150 ^oC, the CLL cells were transferred to a clean 1.5-ml Eppendorf tube and centrifuged at 800 rpm in a bench-top micro-centrifuge for 1 minute. The pelleted cells were washed in 1 ml phosphate buffered saline (PBS) under the same centrifugation conditions before the culture media was removed. For genomic DNA (g. DNA) extraction, Qiagen AllPrepDNA/RNA Mini Kit (Qiagen, UK) was used following the manufacturer's instructions. Briefly, 0.5 – 10 x 10^6 of washed and loosen CLL cells were disrupted completely by slowly adding and mixing by pipetting of appropriate volume (350 - 600 µl) of Buffer RLT (contained 1% of 14.3 M β-Mercaptoethanol (BME)) (Sigma-Aldrich, UK).

The disrupted cells were homogenised by transferring them into a QIAshredder spin column (Qiagen, UK) and centrifuging at 13000 rpm in the micro-centrifuge for 2 minutes. Then, the homogenised lysate was transferred to an AllPrep DNA spin column and centrifuged again at 10000 rpm for 30 seconds. The column bounded DNA was transferred to a new 2-ml collection tube and washed with 500 µl of the provided washing buffer AW1 (diluted with ethanol before use) by spinning at 10000 rpm for 15 seconds. This was followed by the second wash with 500 µl of buffer AW2 (diluted with ethanol before use) and spinning at 10000 rpm for 2 minutes. After one more minute of centrifugation at 14000 rpm for complete dryness, the AllPrep DNA spin column was transferred to a new 1.5-ml collection tube.

At the DNA elution step, 50 µl elution buffer (EB) was added to the column. Following incubation for 1 minute at room temperature, the DNA was spun from the column down to the collection tube at 10000 rpm for a minute. The DNA elution step was repeated again to maximize the yield if necessary. The isolated g. DNA was immediately stored at -20 ^{o}C following the assessment of purity and quantity.

2.2.3. DNA quality assessment

Determining the DNA quality and quantity is essential for downstream molecular biology work. Purity of the isolated DNA from Section 2.2.2 was assessed using a NanoDrop 2000 spectrophotometer (Thermo Scientific, UK). The isolated DNA was considered to be pure if the ratio of UV light absorbance at 260 nm to 280 nm fell within the range of 1.8 - 2.0. A value below 1.8 indicates protein contamination while a value above 2 indicates RNA carry over.

2.2.4. Fluorometric DNA concentration measurements

The DNA concentration was measured using Qubit® 2.0 Fluorometer and Qubit® dsDNA HS Assay Kit (both from Thermo Fisher Scientific, UK). This method is superior to Nanodrop for measuring DNA concentration as the high selectivity to double-stranded DNA (dsDNA) provides more accurate measurement of DNA concentration. According to the manufacturer's manual, the Qubit® working solution was prepared by combining 1 µl of Qubit® dsDNA HS reagent and 199 µl of Qubit® dsDNA HS buffer. In order to produce a standard concentration curve for each batch of experiment, measurement of both the high or low concentration standards provided were performed. The standards were equilibrated to room temperature and then diluted in 190 µl of the prepared working solution using a Qubit® Assay Tube (Thermo Fisher Scientific, UK). For sample DNA measurement, 1 - 4 µl of the sample DNA were diluted in 199 -196 µl of the prepared working solution in each tube. Then each tube was measured on the Qubit® 2.0 Fluorometer. DNA concentration $\geq$ 4.5 ng/µl was considered appropriate for downstream application. For samples with high

concentrations, therefore, a dilution was required to make the DNA concentration 5 ng/µl (range 4.5 - 6.5 ng/µl) with nuclease free water (Thermo Fisher Scientific, UK).

2.2.5. DNA size measurement

DNA size was measured at different steps of this study. This is because firstly the integrity of starting g. DNA is essential for the success of the test, as degraded DNA not only affects efficiency of the test but also the target coverage, and secondly, DNA size reflects outcome of DNA fragmentation, PCR amplification and target DNA purification. The following two approaches were used to measure DNA size.

2.2.5.1. Agarose gel electrophoresis

This technique was repeatedly used in this study to assess results of separation, quantification and purification of DNA fragments based on differential migration for DNA fragments with different number of base pairs (bp). An electrical field separates the fragments where high percentage gels are used for isolating small and low percentage gels for large DNA fragments. The size of separated bands of DNA is determined by comparison to DNA ladders with known molecular weight.

Desired amounts of agarose powder (Sigma-Aldrich, UK) were weighed, then 100 ml of either 0.5 x Tris-borate- Ethyene-diaminen-tetra-acetic acid (EDTA) (TBE) buffer or 1x of Tris-Acetate- EDTA (TAE) buffer was added in a clean autoclaved flask. By using a microwave, the dissolved powder was heated initially for 1 minute, and then for short times until the agarose was completely melted. A few minutes were allowed for the agarose to cool down before pouring it in to a gel tray. With the desired well comb in place, the gel was allowed to set at room temperature for at least 1 hour.

After removal of the comb, the gel was placed in an electrophoretic tank (the loading slots were close to the negative electrode). Then the gel was covered with the same TBE buffer used in gel preparation. After mixing at a ratio of 5:1 with 6 x loading buffer (NEW ENGLAND

BioLabs, UK), separate g. DNA samples and a DNA standard ladder were loaded into the corresponding slots. Following electrode connection to a power pack, the electrophoresis was applied under 100 V constant voltages until sufficient displacement of the tracking dye (bromophenol blue) was observed (for up to 1 hour for that in TBE buffer or up to 16 hours for that in TAE buffer). Then the gel was stained in 200 ml of the corresponding buffer mixed with either 20 µl Ethidium Bromide (EtBr) (10 mg/ ml) (Sigma-Aldrich, UK) or 15 µl of Midori Green Advance DNA stain (GENEFLOW, UK) with continuous shaking for 20 - 40 minutes. DNA bands on the gel were visualised and recorded using an INGenius UV image system (SYNGENE, UK).

2.2.5.2. On chip microfluidic electrophoresis

For a more accurate measurement for DNA digest products and DNA libraries for next generation sequencing, this method was preferable because it is faster, produces a high-quality real-time digital data without necessity of adjustment of electrophoresis time and consumes the least amount of DNA. It was performed using an Agilent 2100 Bioanalyser system and Agilent High Sensitivity DNA Assay Kit (both from Agilent Technologies, UK).

The Bioanalyser system works depending on the passage of separate voltage through 16 micro-channels. Of them, 4 are specified for loading gel dye mixture, 11 for DNA samples mixed with internal DNA upper and lower markers, and the remaining one for DNA size ladder. In this system, smaller DNA fragments migrate faster than the large ones. The signals of fluorescent dyes intercalated into DNA, are then detected and translated into separated DNA band images on gel and DNA peaks and positions in electrophoreogram.

Following the manufacturer's recommendation, the High Sensitivity DNA Dye and the High Sensitivity DNA Gel matrix were allowed to equilibrate from 4 ^{o}C to room temperature for 30 minutes before use. 15 µl of the above dye was added to one vial of the gel matrix provided with the kit and properly mixed using a vortex. This mixture was then transferred to the spin filter and centrifuged at 6000 rpm in a bench-top micro-centrifuge for 15 minutes and stored away from light at 4 ^{o}C for later use.

The High Sensitivity DNA chip was loaded after placing it in to the chip priming station. Initially, 9 µl of the gel-dye mix was loaded into the specified gel loading well using a non-filtered tip and by vertically directing the pipette tip until it touched the bottom of the well. The syringe plunger attached to the prime station was then positioned at the 1-ml mark. With the priming station closed, the plunger was pushed down to the 0-ml mark and clipped by the prime station for 60 seconds to allow the gel to distribute through the interconnected micro-channels of the chip. Then the plunger was released for 5 seconds to allow spontaneous air returning. Next, the plunger was slowly pulled back to the 1-ml mark before the prime station was opened. The same amount of the gel-dye mix was then loaded to each of the other three gel loading wells without creating any pressure. Then, 5 µl of the sample-internal DNA marker were loaded in each of the other 12 sample wells including the ladder well. Following this, 1 µl of the kit provided DNA ladder or 1 µl of DNA samples was loaded into each corresponding well. The loaded chip was vortexed for 1 minute at 2400 rpm on a chip-adapted vortex mixer before placed in the Bioanalyser system. Finally, the Agilent 2100 High Sensitivity assay was performed and results were visualised and analysed using the Agilent 2100 Software.

2.2.6. Gene panel selection

Based on information on CLL recurrent mutations found in WGS and WES studies by others before this study [149, 162, 163, 165], 15 genes were included in the targeted gene panel to be tested for somatic mutations. This was because all of them had an incidence of somatic mutations in $\geq$ 5% of CLL cases (a detail of each gene in the panel is described in Chapter 1, Section 1.2.4). In other words, genes with less frequent mutations were not selected. To increase the coverage depth, and therefore the test sensitivity, only mutated exons were targeted for the sequencing. The details for the targeted exons in the 15 genes are summarised in Table 2.2.

Table 2.2. **Summarised information on the gene panel selected in this study**

Gene name	Accession number (Ensembl)	Chr. location (Ensembl)	Total No. of exons	Targeted exons in this study	Pathway involved	% of affected CLL cases	Mutated exons identified in CLL	References
TP53	ENST00000269305	17p13.1	11	2-11	DNA damage- cell cycle control	15-37%	2-11	[135, 138, 244]
ATM	ENST00000278616	11q22.3	63	2-63		9-14%	2-63	[95, 142]
SF3B1	ENST00000335508	2q33.1	25	7-19	RNA processing	10-17%	12-15	[102, 158]
XPO1	ENST00000401558	2p15	25	11-16		3-5%	11-16	[13, 92]
BIRC3	ENST00000263464	11q22.2	9	2-9	NFκB and inflammatory pathway	4-24%	2-9	[163, 245]
MYD88	ENST00000417037	3p22.2	5	2-5		2-10%	3-5	[168, 169]
SAMHD1	ENST00000262878	20q11.23	16	1-16		7%	1-14	[176]
NOTCH1	ENST00000277541	9q34.3	34	34	NOTCH signalling	5-15%	34	[114, 158]
FBXW7	ENST00000281708	4q31.3	12	5-12		4-5%	5-12	[93]
CHD2	ENST00000394196	15q26.1	39	12-36	Chromatin modification	5%	16-35	[177]
HIST1H1E	ENST00000304218	6p22.2	1	1		5%	1	[198]
POT1	ENST00000357628	7q31.33	19	5-19	Protection of telomeres	4-8%	5-19	[173]
LRP1B	ENST00000389484	2q21.2	91	34-86	Tumour suppression	5%	34-86	[150]
ZFPM2	ENST00000407775	8q23	8	8		5%	8	[246]
PCLO	ENST00000333891	7q11.23-21.3	25	2-24		9%	4-23	[191, 246]

2.2.7. HaloPlex probe design

HaloPlex target enrichment probes were designed using Agilent SureDesign Software (Agilent Technologies, UK) (both Standard and Advanced Wizard settings for optimisation of probe number overlaid GC rich regions. See Sections 2.3.1.3 and 2.3.1.4 for more details). Briefly, the probes were designed for NGS platform (Ion Torrent PGM) with defaulted *H. sapiens* (hg19, GRCh37) as the human reference genome. All DNA targets were selected from the Ensembl archive (http://www.ensembl.org/Homo_sapiens). Genes with full transcripts were selected first to avoid missing any target exons [247]. Then these genes or exons were tailored by the genomic coordinate identifier to subtract non-targeted sequences (See Appendix 7.1 for more details).

To construct the probes, the individual target genomic coordinates were submitted to the SureDesign Software. In addition, 10 bp from both 5' and 3' ends for each target region were extended outside of each target sequence to ensure coverage of splice sites junctions, as mismatches in these sites tend to have effects on gene expression by affecting the binding with the spliceosome for proper mRNA splicing [248]. The target regions and the amplicons covering the target sequences were viewed using USCS Genome Browser (www.http://genome.ucsc.edu).

In total, 4606 amplicons that covered 135.420 Kbp of sequenceable region, including 52.324 Kbp of targets were designed. The total sequenceable size and the average output of Ion 318 Chip were used to estimate the average depth of coverage (the number of sequencing reads of each target base) for each sample sequenced.

2.2.8. Target enrichment using HaloPlex technique

Following the quantity and quality assessments (described in Sections 2.2.3 to 2.2.5), ≥ 40 µl of DNA at 4.5 - 6.5 ng/µl were prepared from each sample for target enrichment using the HaloPlex technique.

2.2.8.1. Shearing of genomic DNA

HaloPlex Target Enrichment System (Agilent Technologies, UK) starts with enzymatic digestion of the genomic DNA. This creates fragments sized between 100 - 500 bp to accommodate the read length readable by Ion Torrent PGM. Furthermore, it forms complimentary sites critical for hybridisation of the designed probes. 16 different restriction enzymes were provided in two 8-well strips marked in red and green, respectively. This allowed digesting of each sample DNA in eight combinations.

In a 1.5-ml Eppendorf tube, 34 µl of restriction enzyme buffer (RE) and 0.85 µl of bovine serum albumin (BSA) solution, functioning as a stabilizer of the restriction enzymes, were mixed for each sample digestion reaction. Then, 4 µl of the RE-BSA mixture was distributed into each of the 8 (0.2-ml) digestion tubes on ice for individual digestion enzyme combination. For each sample reaction, 0.5 µl of restriction enzyme from well A - H of the red strip was added respectively to the digestion tubes 1 - 8. In the same way, enzymes from the green strip were added to the 8 digestion tubes. This was followed by gentle vortexing and a brief spinning. Then on ice, 5 µl of the prepared g. DNA from one sample was added to each of 8 digestion tubes. After a gentle vortex and a brief spin, the digestion tubes were incubated at 37 ^{o}C for 30 - 55 minutes in a thermo-cycler. The temperature of the preheated lid was set at 45 ^{o}C to prevent evaporation. The reaction was then terminated by transferring these tubes to a -20 ^{o}C freezer for further processing next day.

2.2.8.2. Controls for g. DNA digestion

In 8 different 0.2-ml tubes, 4 µl of each the 8 digestion reactions from the CLL samples were heated at 80 ^{o}C for 5 minutes to inactivate the restriction enzyme. 1 µl of each heat inactivated restriction reaction and corresponding undigested DNA as control were loaded on a Agilent High Sensitivity DNA Assay chip using Agilent Bioanalyser 2100 for assessing results of the digestion as described in Section 2.2.5.2.

2.2.8.3. Hybridisation of g. DNA to HaloPlex probe and sample barcoding

In this step, the designed target specific probes selectively hybridised to target regions of the genome. These single strand probes with a length of 100 bp, including index bound to both ends of the target DNA fragments to form nicked DNA circles which facilitate subsequent ligation. Up to 16 molecular barcodes were also incorporated into each batch of DNA samples. They were later used for identification of specific samples in the data analysis of pooled DNA sequencing.

For each DNA sample, 47 - 50 µl of supplied Hybridisation Solution and 17 - 20 µl of the HaloPlex Probes were mixed in a 0.2-ml hybridisation tube. After a brief vortex and spin, 10 µl of a molecular barcode was added to the sample tube. Then the 80 µl of digested DNA (Section 2.2.8.1) were transferred individually from the 8 digestion tubes to the hybridisation tube and carefully pipetted up and down for at least 10 times. This was to properly mix DNA fragments with the probes and the barcode and completely inactivate digestion enzymes in the hybridisation buffer. Following a brief spin, the hybridisation tube was transferred into a thermal cycler block with temperature of the preheated lid set above 95^{o}C and the hybridisation started under conditions presented in Table 2.3.

Table 2.3. **Conditions for hybridisation of g. DNA with HaloPlex probes and for barcoding**

Step	Block temperature	Time
1	95 ^{o}C	10 min.
2	54 ^{o}C	3 hr.

In the meantime, a required volume of HaloPlex beads (40 µl/ sample) for the next process were transferred from the 4^{o}C storage to a 0.2-ml or 1.5-ml tube to allow the bead temperature to restore to room temperature within 30 minutes. The bead tube was then

placed on a magnetic rack for 5 minutes to allow the beads to sink down before the supernatant was removed. Then, the same volume of the provided Capture Solution was added to re-suspend the beads in the tube which had been separated from the magnetic rack.

After completion of hybridisation reaction, 40 µl of the homogenised HaloPlex bead suspension was added immediately to the hybridisation tube and mixed by pipetting up and down 15 times. The tube was then left at room temperature for 15 minutes.

Following the 15 minutes incubation, the tube was briefly centrifuged and placed on the magnetic rack until the solution became clear. Then, the supernatant was discarded using a pipette set to 200 µl. By this step, the entire off-target and non-hybridised DNA was isolated from the target region. The capture reaction was removed from the magnetic rack and 100 µl of Wash Solution was added and mixed by pipetting up and down for 10 times. This tube was then incubated at 46 ^{o}C for 10 minutes in a thermal cycler block with preheated lid. Immediately following the completion of incubation, the tube was briefly spun and re-placed on the magnetic rack until the solution became clear. The supernatant was carefully removed using a pipette set to 120 µl firstly and another at 20 µl.

2.2.8.4. Ligation of the captured (circularised) DNA fragments

In this step, the nicks of the circularised HaloPlex probe bounded target DNA are closed by using a DNA ligase. The ligation master mix was prepared by combining 47.5 µl of Ligation Solution and 2.5 µl of DNA Ligase for each sample reaction. After vortexing and brief spinning, 50 µl of the ligation master mix was added to a tube containing captured DNA and mixed by pipetting up to 15 times. The reaction tube was incubated at 55 ^{o}C for 10 minutes using a thermal cycler programme set with a heated lid. Then, the tube was taken out from the thermal cycler and briefly centrifuged before transferred to the magnetic rack. After the solution became clear, the supernatant was carefully discarded with a pipette set to 50 µl. Next, the tube was removed from the magnetic rack and then loaded with 100 µl of provided Saline Sodium Citrate (SSC) Buffer to re-suspend the beads by pipetting up to 10 times. Following a brief spin, the tube was placed into the magnetic rack, and the clear SSC supernatant was removed with a pipette set at 120 µl.

The final elution step of the bead bound DNA was performed by adding 25 µl of freshly prepared 50 mM NaOH (Sigma-Aldrich, UK) to the tube. After mixing by carefully pipetting up to 10 times, the sample was incubated at room temperature for 1 - 2 minutes. Then, the tube was centrifuged briefly and transferred to the magnetic rack. Probe-bound DNA in 20 µl of the cleared supernatant was collected and used in the next step for PCR amplification.

2.2.8.5. PCR amplification of DNA libraries

PCR master mixture was prepared and kept on ice by combining the provided PCR primers and other reagents in a 0.2-ml tube as presented in Table 2.4.

Table 2.4. **PCR components for HaloPlex captured DNA library amplification**

Supplier of reagents not provided with HaloPlex kit	**Reagent**	**Amount for 1 reaction** (µl)
(Agilent Technologies, UK)	5x Herculase II Reaction Buffer	10
	dNTP mix (100 mM)	0.4
	HaloPlex Ion primer 1 (25 µM)	1.0
	HaloPlex Ion primer 2 (25 µM)	1.0
(BD Biosciences, UK)	2 M Acetic acid	0.5
(Agilent Technologies, UK)	Herculase II Fusion DNA polymerase	1.0
(Sigma-Aldrich, UK)	Nuclease-free water	16.1
	Total	30

The 20 µl of eluted target DNA was added to the above PCR mixture and PCR was performed in a lid-heated thermal cycler (Mastercycler AG Eppendorf, Germany) with the following programme shown in Table 2.5.

Table 2.5. **Programme temperature settings for PCR amplification of HaloPlex captured target DNA libraries**

Step	No. of cycles	Temperature (°C)	Time
1	1	98	2 min.
2	22	98	30 sec.
		60	30 sec.
		72	1 min.
3	1	72	10 min.
4	1	8	Hold

The PCR product (amplified DNA library) was used in next step either immediately or after overnight storage at -20 °C.

2.2.8.6. Size selection

In this step, Agencourt®AMPure® XP beads (Beckman Coulter, UK) were used to purify the amplified DNA libraries from any excess probes and primers.

Before starting, the XP beads were equilibrated to room temperature for 30 minutes. The bead suspension was vigorously vortexed until the suspension became consistent in colour. For each sample to be purified, 50 - 100 µl of homogenous beads was mixed with 20 - 40 µl of nuclease free water in a 0.2-ml PCR tube by vortexing. Accordingly, either 70 or 140 µl of diluted bead was prepared to purify 20 or 40 µl of the amplified library DNA, respectively. The 90 - 180 µl sample-bead mixture was incubated for 5 - 15 minutes (at room temperature) with continuous shaking using a shaker. The tube was then briefly centrifuged and placed in 0.2 ml tube-adapted magnetic rack for 5 minutes to clear the solution. Using a pipette set at 100 or 200 µl, the supernatant was carefully removed. Any residuals were discarded by a pipette set at 10 µl. While the tube was retained in the magnetic rack, an ethanol washing step was performed twice by slowly adding the appropriate volume (100 - 200 µl for the sample-bead mixture of 90 µl -180 µl, respectively) of freshly prepared 70% molecular grade ethanol (Sigma-Aldrich, UK). After each 30 seconds wash, the supernatant

was removed with a pipette. The residual ethanol was carefully removed with a 20 µl tip and the beads were left in the tube with lid open for air drying for 5 minutes. The dried bead tube was removed from the rack and the beads were mixed with 20 - 40 µl of 10 mM Tris-HCl buffer (Sigma-Aldrich, UK) by pipetting up and down for 15 times.

To allow complete elution, the tube was incubated at room temperature for 2 minutes and then placed in the magnetic rack until the solution cleared. The supernatant (20 - 40 µl) was then transferred in to a fresh 0.2-ml PCR tube for repeating AMPure XP bead purification and then stored at -20 ^{o}C.

2.2.9. Validation of DNA library amplification

This verification step is essential before proceeding to the next step not only for validating efficiency of the library amplification, but also for determining purity and quality of the resulting library. For this purpose, 1 µl of each purified DNA library was loaded on an Agilent High Sensitivity DNA Assay chip using Agilent Bioanalyser 2100 as in Section 2.2.5.2. Successfully amplified and purified DNA library appeared as DNA fragments sized from 150 to 550 bp, and did not contain readily visible free probes and PCR primers (35 - 100 bp).

2.2.10. Nanomolar concentration measurements, pooling of DNA libraries and calculating the dilution factor

Measurement of DNA library concentration in (ng/µl) was achieved by using Qubit® dsDNA HS Assay Kit as in (Section 2.2.4). The average size (in bp) of DNA fragments as measured with the Bioanalyser, was used to calculate approximate molecular weight (MW) for double stranded (ds) DNA based on the following: the average molecular weight of the four di-nucleotides: G, C, A and T being 303.7 and the molecular weight of triphosphate molecule from the end of each DNA strand being 78.95. The following equations were used in this calculation

1). MW of dsDNA = (Average number of nucleotides x 607.4) + 157.9
2). Nanomolar (nM) concentration of the DNA library = DNA library concentration (ng/ µl) ÷ MW of ds DNA x 1000

Nanomolar (nM) library concentration was used for calculating the dilution factor to obtain the desired DNA molecules per 5 µl of diluted DNA library. 70 million - 140 million DNA molecules are required to obtain maximum number of mono-clonally enriched ISPs with lowest polyclonal enrichment. To quantitatively equalise the multiple (4 - 5) DNA libraries co-sequenced on an Ion chip, the concentration of individual libraries were adjusted with dilution before sample pooling. The pooled DNA libraries finally reached at 10 - 25 pM in no less than 25 µl. They were then vortexed and kept on ice until used.

2.2.11. Preparation of template- positive Ion Sphere Particles

At the end of this step, each solid spherical particle is expected to be attached by a single DNA library molecule. Through emulsion PCR, the DNA library clonal amplification occur on the surface of the ISPs which then by loading and directional centrifugation of the loaded chip, are dislodged into the micro-wells.

For this purpose, initially Ion OneTouch™ 200 Template Kit v2 DL (Life Technologies, UK) and Ion OneTouch™ Instrument (Life Technologies, UK) were used. Later, due to discontinuation in market, they were replaced by Ion PGM™ Template OT2 200 Kit (Life Technologies, UK) and its corresponding compatible instrument Ion OneTouch™ 2 (Life Technologies). The procedural steps of the two systems were similar, except the new system seemed to perform better and was more cost effective due to eliminated request for Argon gas source.

According to the manufacturer's manual, the whole procedures including DNA library amplification, recovery of the enriched ISP and annealing of the sequencing primers were carried out consecutively. Detailed experimental steps are found in the Appendix 7.3.1.

2.2.12. Sequencing on the Ion Torrent PGM

All reagents were from the Ion PGM™ Sequencing 200 v2 Kit (Life Technologies, UK) and Ion 318™ Chip Kit v2 (Life Technologies, UK). Cleaning, initialising and chip checking steps were performed before loading of the sequencing chip with the template positive ISP.

Automated cleaning of the Ion Torrent PGM machine (Life Technologies, UK) was performed with either 18 MΩ water alone or with Chlorite followed by 18 MΩ water washing. Detailed information on each step conducted can be found in the Appendices 7.3.2, 7.3.3 and 7.3.4.

2.2.13. Assessment of the sequencing runs

The quality of a run is dependent on many factors including quality and quantity of the template library and chip loading density. In each sequence run, the server provided summary statistics reports in accordance to the chip type and the sequencing kit used in the experiments.

2.2.14. Data collection, processing and reporting

Torrent Suite Pipeline v4 (Life Technologies, UK) was used for raw sequencing data processing. The output files were sorted by the HaloPlex Barcode specific for individual samples combined in each run. Total sequenced bases, the number of bases with Q scores >20, the number of reads and the mean read lengths were then calculated. Compared with the human reference genome hg19, the numbers of aligned reads were determined. The whole customised library target region BED file obtained from the SureDesign (HaloPlex) software was then uploaded for analysis with Torrent Server Reference and Target Region tags. In the same way, an individual target region BED file for each gene was created for assessing the coverage of each target region by Coverage Analysis Plugins (v4.4.2.2). A total of 544 coverage analyses were performed to identify average base coverage depth, uniformity of coverage and percentage of base coverage for the DNA targets studied.

Torrent Variant Caller (TVC) Plugin (v4.2.1.0) was used to call somatic variants. With the hg19 as reference, analysis was performed with or without confinement to target and hot spot regions. The generic PGM analysis stringency settings were tested and customised

based on known levels of multiple mutations in diluted patient samples. Through adjusting various parameters including minimum frequency of variant alleles, minimum quality and coverage, and maximum strand bias, the customised settings should maximise the specificity and precision of analysis. After completion of TVC runs, a list of variants' chromosomal positions, reference and variant bases, allele frequencies, with quality and coverage metrics were produced for each sample. After visualising each called variant by Integrative Genomic Viewer (IGV), the output VCF files were uploaded to Ion Reporter Software v5.0 (Thermo Fischer Scientific, UK) using the user specific account. After defining the imported sample files, the analysis work flow (set to Annotate Variants) was launched. The results were viewed for details of gene ID, variant location, c. DNA and protein change, gene ontology and functional consequences for each variant. The software integration with catalogue of somatic mutations in cancer (COSMIC-65) (http://www.sanger.ac.uk/cosmic), single nucleotide polymorphism (dbSNP-137) (http://www.ncbi.nlm.nih.gov/SNP) and 1000 genome project (http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/) databases helped in determining the type of each variant.

With the above information about variants in each sample, a master data file was produced in Excel 2010 spreadsheet. The data sheet contained data on gene ID, chromosomal position, reference allele, variant allele, c. DNA position, codon and amino acid change. Moreover, information from publicly available databases including dbSNP ID, COSMIC ID, and from in silico bio-informatics tools including sorting tolerant from intolerant (SIFT) (http://sift-dna.org), polymorphism phenotyping (PolyPhen-2) (http://genetics.bwh.harvard.edu/pph2/) scores were recorded. Furthermore, allele frequency, quality and coverage metrics for all the identified variants in the sequenced samples in this study were included in this datasheet. This enabled multiple samples to be compared and shared variants between different samples to be identified. Moreover, replicates of sequencing experiments of the same case and sequential samples of individual patients were compared and tracked using the same spreadsheet.

For the purpose of identifying somatic non-synonymous mutations, variant calls outside coding regions of targeted genes were firstly filtered out. The remaining variants were then compared to those reported in dbSNP-137 and 1000 Genome Browser to further filter out

variants that were population-study confirmed germline polymorphisms. Next, the candidate somatic non-synonymous variants were visually checked in IGV display. Finally, the amino acid changes resulted from these somatic non-synonymous variants and their biological effects were determined and estimated by comparison with public databases including COSMIC-65, SIFT and PolyPhen-2.

2.2.15. Allele specific PCR (AS-PCR) for validation of low level mutations

AS-PCR was used to test the fidelity of mutations detected by Ion Torrent PGM in a sensitivity test experiment in which different *TP53* mutations were diluted to 20% - 0.2% variant allele frequency (VAF). Primers were designed based on human *TP53* sequence (NC-000017.11, NCBI) and using a common tool (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). The last nucleotide at the 3' end of the reverse primers were complimentary to either wild type (Wt) or mutant (Mu) allele, while a consensus forward primer was designed for both wild-type and mutant alleles (Figure 2.1). All primers were ordered from Eurofins Genomics, Germany.

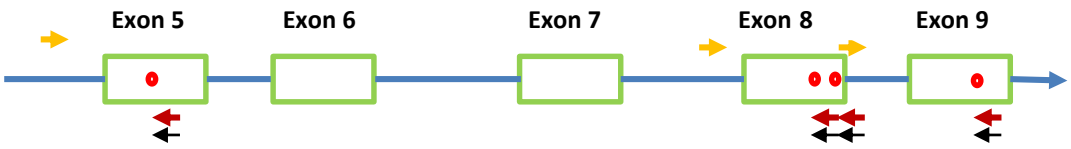


Figure 2.1. **Locations of low level *TP53* mutations and the primers used for As-PCR** The 3' end nucleotide of each reverse primers was complimentary to either Wt (black arrows) or Mu alleles (red arrows) for each mutation (red dots), while consensus forward primers (orange arrows) were used for both wild-type and mutant alleles amplifications.

Qiagen Taq polymerase kit was used and the amplification reactions were conducted under optimised conditions using the specific primers designed as summarised in Table 2.6. The products of AS-PCR were then examined by electrophoresis. For this purpose, 10 µl of each PCR product was mixed with 2 µl of 6 x loading dye and loaded along a 100 bp DNA size marker on 1% agarose gel. The gel was run at a constant 100 V in 0.5 x TBE for 90 minutes. The gel was then visually examined under UV illumination after staining with EtBr as described in Section 2.2.5.1.

2.2.16. Statistics

In this Chapter, the tendency of centralisation and dispersion of data distribution was presented as mean $\pm$ standard deviation (SD), or mean $\pm$ standard error (SE) if used for multi groups which had unequal number of samples. Pearson correlation was used to investigate relations between two variables, with the α level set to 0.05 for two side test. The IBM SPSS (version 20) was employed for all statistical analyses.

Table2.6. **The allele specific PCR information on variants, primer sequences and corresponding PCR condition for confirmation of 4 variants in *TP53* gene**

Location and nucleotide change (ref. hg19)	Codon changes	Amino acid changes	Primer names	Primer sequence (5'-3') NCBI Reference Sequence Accession Number (NC-000017.11)	PCR components 50 µl in 0.2-ml thin walled PCR tubes								Amplification condition *			Size of PCR products (bp)
					Primer (Pmole)		$MgCl_2$ (mM)	dNTP (mM)	GoTaq (U)	g.DNA (ng)	5x buffer (µl)	DMSO (µl)	Temp. (^{0}C)	Time (Sec.)	No. of cycles	
					Wt.	Mu.										
17:7577079 C>A	**G**AG>**T**AG	p.287 E/X	287G-Rev (Wt) 287T-Rev (Mu) 280S-For	CCCTTTCTTGCGGAGATTCT**C** CCCCTTTCTTGCGGAGATTCT**A** TACCCATCCACCTCTCATCAC	20 ----- 20	----- 20 20	75	10	1.25	84	10	------	94 59 72	30 30 30	32	338
17:7576891 T>A	**A**AG>**T**AG	p.319 K/X	319A-Rev (Wt) 319T-Rev (Mu) 280S-For	CTCCATCCAGTGGTTTCTTCT**T** CTCCATCCAGTGGTTTCTTCT**A** TACCCATCCACCTCTCATCAC	20 ----- 20	----- 20 20	75	10	1.25	84	10	-------	94 59-60 72	30 30 30	32	527
17:7578394 T>C	C**A**T>C**G**T	p.179 H>R	179A-Rev (Wt) 179G-Rev (Mu) 179S-For	CGCTATCTGAGCAGCGCTCA**T** CGCTATCTGAGCAGCGCTCA**C** AAAGCTCCTGAGGTGTAGACG	10-20 ----- 10-20	----- 10-20 10-20	25-75	2.5-10	1.25	60	10-20	+/- (1-2)	94 59-67 72	30 30 30	30-32	334
17:7577100 T>C	**A**GA>**G**GA	P.280 R>G	280A-Rev (Wt) 280G-Rev (Mu) 280L-For	CCTCTGTGCGCCGGTCTC C**T** CCTCTGTGCGCCGGTCTC C**C** GAAGACTCCAGGTCAGGAGC	20 ----- 20	----- 20 20	25-75	7.5-10	1.25	70	10	------	94 59-66 72	30 30 30	32	427

*The initial denaturation temperature of 94 ^{o}C for 3 minutes and the final extension temperature of 72 ^{o}C for 5 minutes were used in common for all the PCR reactions

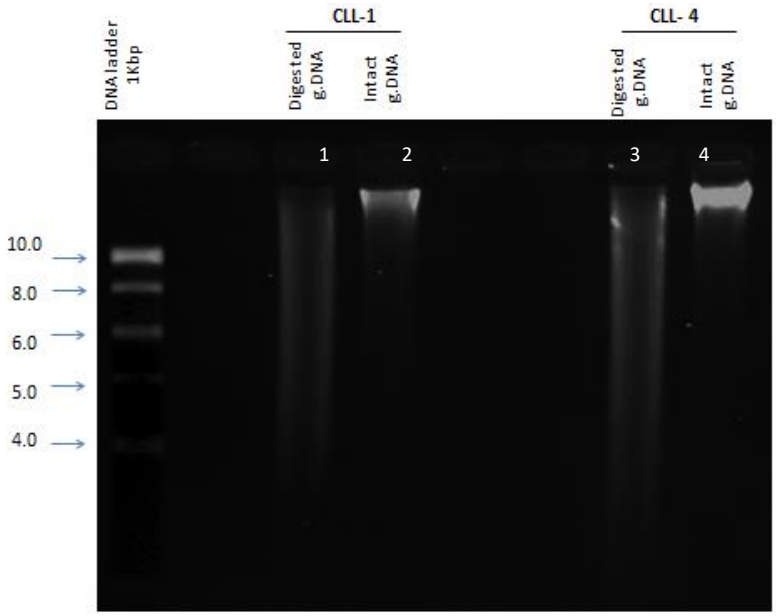
2.3. Results

2.3.1. Modification and optimisation of test conditions

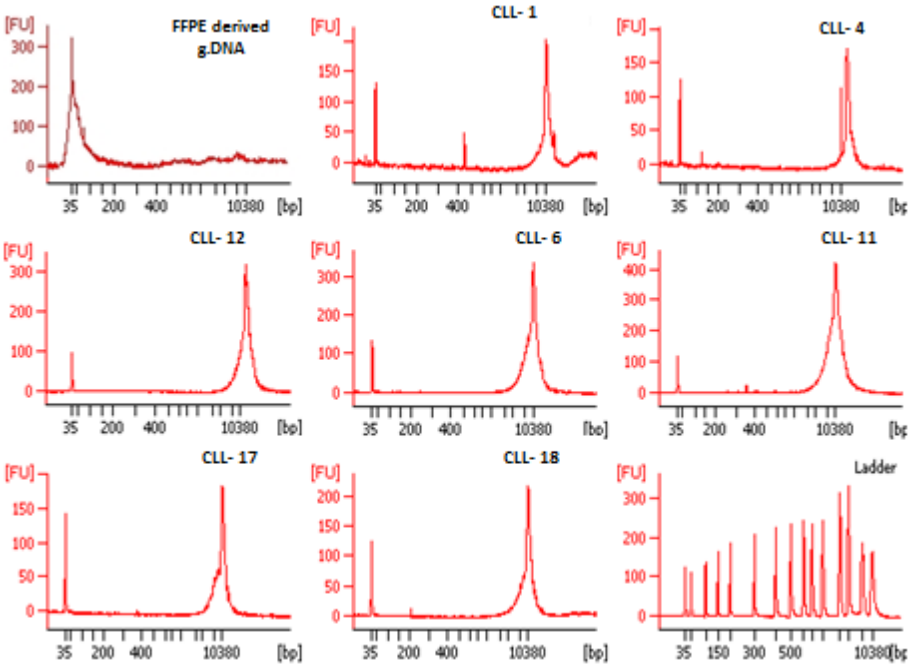
2.3.1.1. Good integrity of starting g. DNA

To avoid missing targets, it is important to start the NGS with intact g. DNA. For this reason, the integrity of extracted g. DNA from each sample was examined before fragmentation. In fact, starting g. DNA extracted from all samples used in this study were intact, appearing as a main band of > 10000 bp as measured with agarose gel electrophoresis as shown as an example in lanes 2 and 4 of Figure 2.2.A, or with Bioanalyser as shown in Figure 2.2.B. This was further verified by amplification of a house-keeping gene fragment of different sizes. Compared to partially degraded g. DNA from a FFPE sample, a large fragment (980 bp) of *GAPDH* was readily amplified using g. DNA extracted from CLL samples for this study as shown as an example in Figure 2.2.C and D.

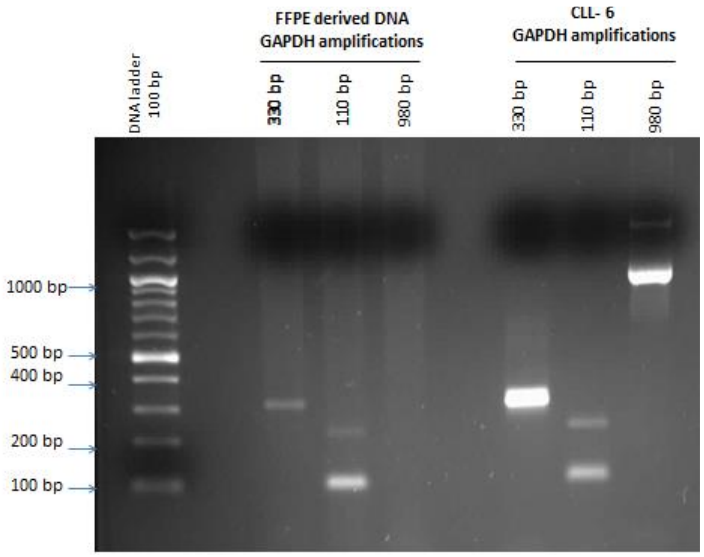
A



B


FFPE derived g.DNA
CLL- 1
CLL- 4
CLL- 12
CLL- 6
CLL- 11
CLL- 17
CLL- 18
Ladder
[FU]
[bp]
35 200 400 10380
35 150 300 500

C


DNA ladder 100 bp
FFPE derived DNA GAPDH amplifications
CLL- 6 GAPDH amplifications
330 bp
110 bp
980 bp
1000 bp
500 bp
400 bp
200 bp
100 bp

D

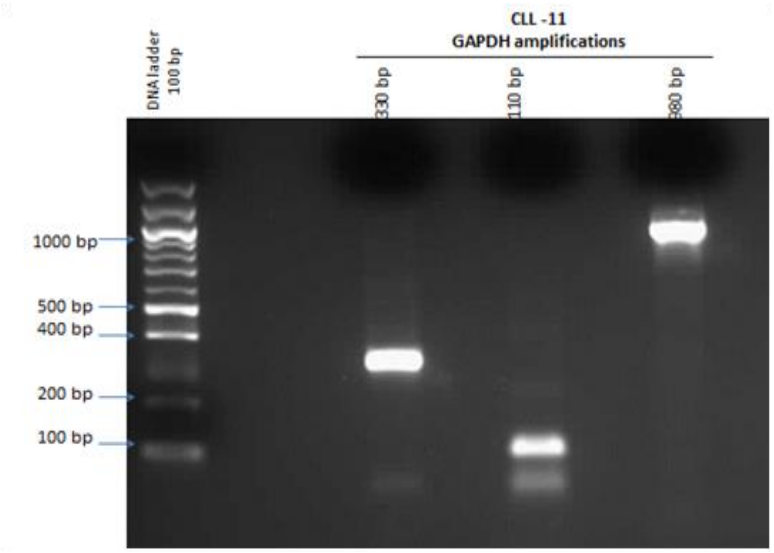


Figure 2.2. **High integrity of starting g. DNA**

As an example, 1.5 mg g. DNA extracted from two cell samples (CLL-1 and CLL-4) were examined by electrophoresis on agarose gel before (lanes 2 and 4) and after (lanes 1 and 3) digestion with EcoR1 (**A**). The undigested, but partially degraded g. DNA prepared from a lacrimal sac FFPE sample and g.DNA from CLL cells of CLL-1, 4, 12, 6, 11, 17 and 18 were visualised in Bioanalyser electrophoreogram (**B**). Electrophoresis on agarose gel showing *GAPDH* fragments of 110, 330 and 980 bp amplified in PCR from the intact g. DNA of CLL- 6 (**C**) and CLL-11 (**D**) and partially degraded g. DNA prepared from the lacrimal sac FFPE sample (**C**).

2.3.1.2. Optimisation of digestion time for g. DNA fragmentation

The length of DNA fragments is equally important as the integrity of the starting DNA. Fragments that are too big or too small affect the efficiency of target enrichment, amplification and/or sequencing. For Ion Torrent PGM, the ideal size of these DNA fragments is between 100 and 500 bp and these are determined by multiple factors, including the restriction map of digestion enzymes used and the limit of sequencing read length that the platform reaches. Conditions for digestion are keys for controlling the size of DNA fragments. Since most of them are fixed, we could optimise the condition only by adjusting durations of the digestion. Experiments digesting a CLL g. DNA sample for 4 different durations, including

the recommended 30 minutes found that the best result was from the digestion for 55 minutes, as the under-digestion was evident with the other time points tested, as shown in Figure 2.3.

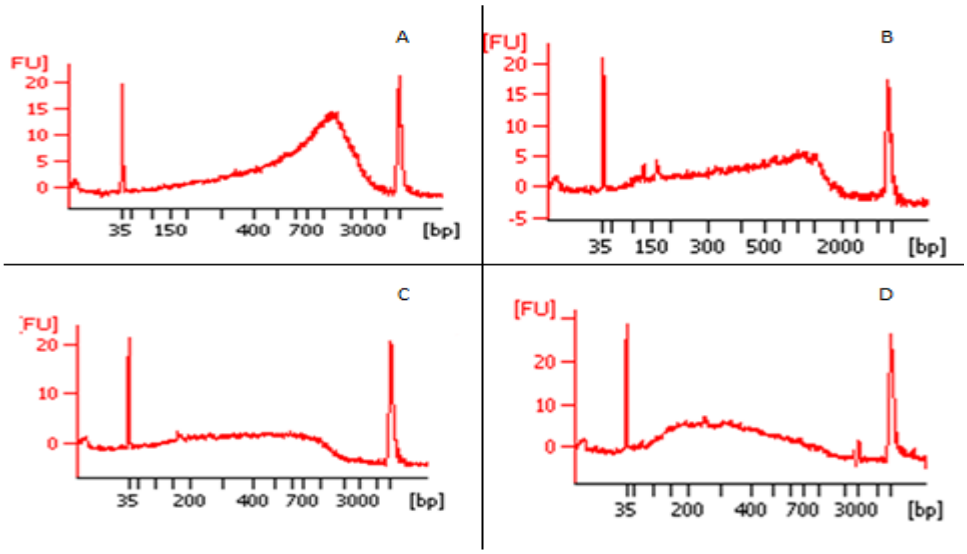


Figure 2.3. **Comparison of g. DNA fragment size resulted from different digestion durations**
G. DNA from a CLL sample (CLL-1) digested with HaloPlex restriction enzymes for **A.** 20 minutes, **B.** 30 minutes, **C.** 45 minutes and **D**. 55 minutes was examined with Bioanalyser electrophoreogram. The curves of fluorescent intensity between two peaks (DNA size markers) present the average and range of digested g. DNA.

2.3.1.3. HaloPlex probe design and coverage depth achievable

As shown in Table 2.7, there were a total of 246 target regions in the 15 genes with a total size of 52324 bp for this study using HaloPlex Standard Design Wizard (Agilent Technologies, UK). Since each target read was covered by multiple probes which could extend to adjacent regions, the number of amplicons was 4585 and the sequenceable size was 135420 bp. Although the software recognised all the input genomic coordinates and estimated the coverage for all the targets to 100%, the target regions were not covered uniformly. Thus some regions, particularly those with AT- or GC-rich sequences, had less chance to be

covered. To improve the coverage of a GC-rich region in exon 4 of *TP53* gene (75 bp in length), HaloPlex Advanced Design Wizard was used and 19 additional probes (amplicons) were created. The software recognised these regions and reported them as 3 additional regions. Therefore, the total number of targets, amplicons and sequenceable region increased to 249, 4604 and 136425 bp, respectively. However, the additional targets and the sequenceable regions still located within the previous target region and therefore should not be counted as increased width.

Table 2.7. **Information about two designs for HaloPlex probes**

Variables	Standard	Advanced
Number of genes	15	15
Number of target regions	246	249
Number of amplicons	4585	4604
Target region size (bp)	52324	52475
Total sequenceable design size (bp)	135420	136425

In order to have the test sensitive enough to detect as low as 1% variant alleles in patients' samples, we planned to achieve an average target coverage depth to 2000 x with the minimum number of variant reads being 20. Based on our design, the estimated average coverage depth for each targeted base was calculated as follows:

Average coverage = average sequencing output of Ion chip /total sequenceable design size x number of samples = 1 Gbp / 135.42 Kbp x 4 = 1846 x

Given that the coverage for adjacent regions is smaller, the depth for the targets is bigger than the figure calculated. To keep a balance between test sensitivity and cost, we planned to include 4 samples on a chip for sequencing with an expected average coverage depth for targets of around 2000 x.

To test if the goal could be achieved, 24 DNA libraries prepared from the 11 local CLL patients and 5 from the ERIC study (Section 2.2.1) were prepared and sequenced on 6 chips. The sequencing quantity and quality are presented in Table 2.8. All experiments were conducted using the successive target enrichment and sequencing approaches described in Sections 2.2.9 - 2.2.16. The results for all the experiments showed highly specific target enrichment as shown in an example in Figure 2.4. Moreover, the produced sequencing data were of high quality and the estimated coverage depth was approached in each experiment. As expected, the average on-target coverage reached to 2198 $\pm$ 192 (Mean $\pm$ SD) when 4 DNA samples were sequenced on each Ion Chip318, while it reduced to 1789 $\pm$ 259 when 5 samples were on each chip. These results clearly showed that our plan was achievable which allowed us to keep the total cost for enrichment and sequencing each sample around £265.

2.3.1.4. Improvement of uniformity and target coverage in a GC-rich region with boosted HaloPlex probe design

Although a satisfactory average uniformity and coverage depth were achieved as shown above, some target regions might not be properly covered. This is because the high throughput sequencing technique cannot produce absolute uniformity for all the targeted regions mainly due to variability in nucleotide composition. GC- or AT-rich target regions, as defined by presence of > 55% GC or AT in the sequence composition, frequently hamper the efficiency of the target amplification by PCR [249]. When examined all of the 246 target regions using IGV, we found a 75-nucleotide GC-rich (73%) region in exon 4 of *TP53* (chr17:7579400-7579475) that was not covered at all.

To overcome this problem, we applied the advanced setting in design to boost probes and enhance their overlay over the GC-rich region. As expected, this approach indeed significantly improved the uniformity (from 73.41% to 92.82%) and target coverage (78.28% to 100%, for $\geq$ 100 x) for that target region (Table 2.9). These improvements were predictable at the time of the advanced probe design in the BED file and clearly visualised after target enrichment and sequencing as shown by IGV (Figure 2.5).

Table 2.8. **The sequencing performance and the average depth of coverage of 6 sequencing runs used in production of data in this chapter** *

Parameters		Run-1	Run-2	Run-3	Run-4	Run-5	Run-6	Mean ±SD
Number of DNA libraries loaded		4	4	4	5	4	4	
Sample's ID		CLL-1,4&24	CLL-1,4,6 &24	S-20,5,1&0.2	ERIC-1,2,3,4&5	CLL-11,12,17&18	CLL-15, 22, 30 & 32	
% of ISP loaded wells		78	82	63	64	63	77	71±8.7
Output (Mbp)		815	827	725	735	792	899	798.83±64.28
% of polyclonal reads		13	12	24	19	18	28	19±6.1
Total useful reads		5981418	6000875	5073100	4946051	5379471	5587162	5494679±44581
Av. coverage depth on targets		2200	2466	1978	2008**	2173	2368	2198.83±192.79
Av. No. of bases with AQ17 (Mbp)/chip		729	784	636	653	705	818	720.83±71.47(180.20±17.8/sample)
Av. No. of bases with AQ20 (Mbp)/chip		665	718	572	618	646	762	663.5±68.44(165.87±17.2/sample)
Median read length (bp)		137	141	142	146	144	161	145.16±8.32
Av. uniformity of coverage		91.1	92.56	92.40	92.43	93.31	92.37	92.36±0.7
Av. % target base coverage at	1x	99.46	99.88	99.59	99.75	99.68	99.49	99.64±0.16
	20x	99.31	99.83	98.69	98.80	98.85	98.85	99.05±0.43
	100x	98.20	97.32	97.19	97.10	97.68	97.68	97.52±0.4
Av. % of reads on target		89.29	93.47	90.88	90.85	89.45	89.45	90.56±1.59
Sequencing kit used		One Touch	One Touch	One Touch	One Touch 2	One Touch 2	One Touch 2	

*In Run-1, CLL-1 & CLL-6 marked in bold had 2 DNA preparations loaded on the same chip. CLL1, 4 and 24 from Run-1 were subjected to DNA preparation with probe boost for GC rich regions and re-sequenced on another chip in Run-2.

** The figure is equivalent to that for 4 DNA libraries sequenced on an Ion chip.

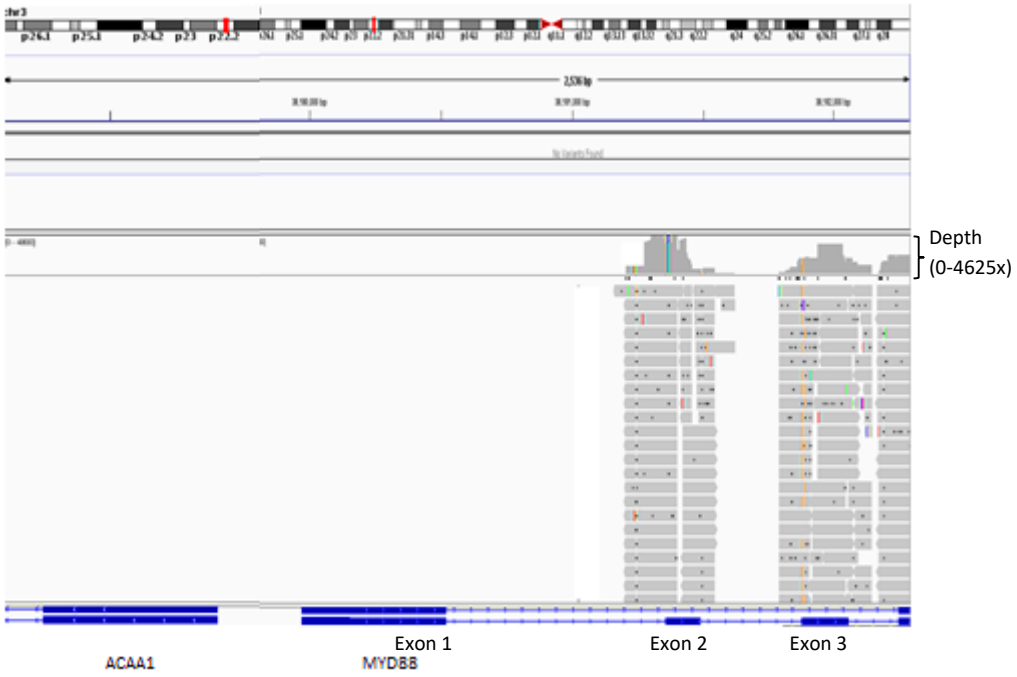


Figure 2.4. **An example of high specificity of HaloPlex target enrichment**

A part of sequence reads (horizontal grey bars) covering the targeted exons 2 and 3 and part of exon 4 (solid blue bars) of *MD88* gene were shown together with the coverage depth (vertical grey or colour bars above the reads). There was no off-target coverage on exon 1 of this gene and an exon of the adjacent gene *ACAA1*, as well as intron-only regions.

Table 2.9. **Comparison of performance of two sets of HaloPlex probe in 3 CLL samples sequenced with or without probe enhancement for a GC-rich target region within exon 4 of *TP53***

Parameters		Sequencing results without boosted HaloPlex probe	Sequencing results with boosted HaloPlex probe
% of Ion 318 Chip loading		78	82
Av. coverage depth of the GC rich target gene		2208 x	2221 x
% of coverage uniformity in the GC rich target gene		73.41	92.82
% of base coverage of GC rich target gene at	1x	91.16	100
	20x	91.16	100
	100x	78.28	100

*Targets from g.DNA of CLL patients (n=3) were enriched with HaloPlex probes without (middle column, Run-1 in 2.5.1.) and with (Right column, Run- 2 in 2.5.1.) boosted probe design and then sequenced using Ion Chip 318.

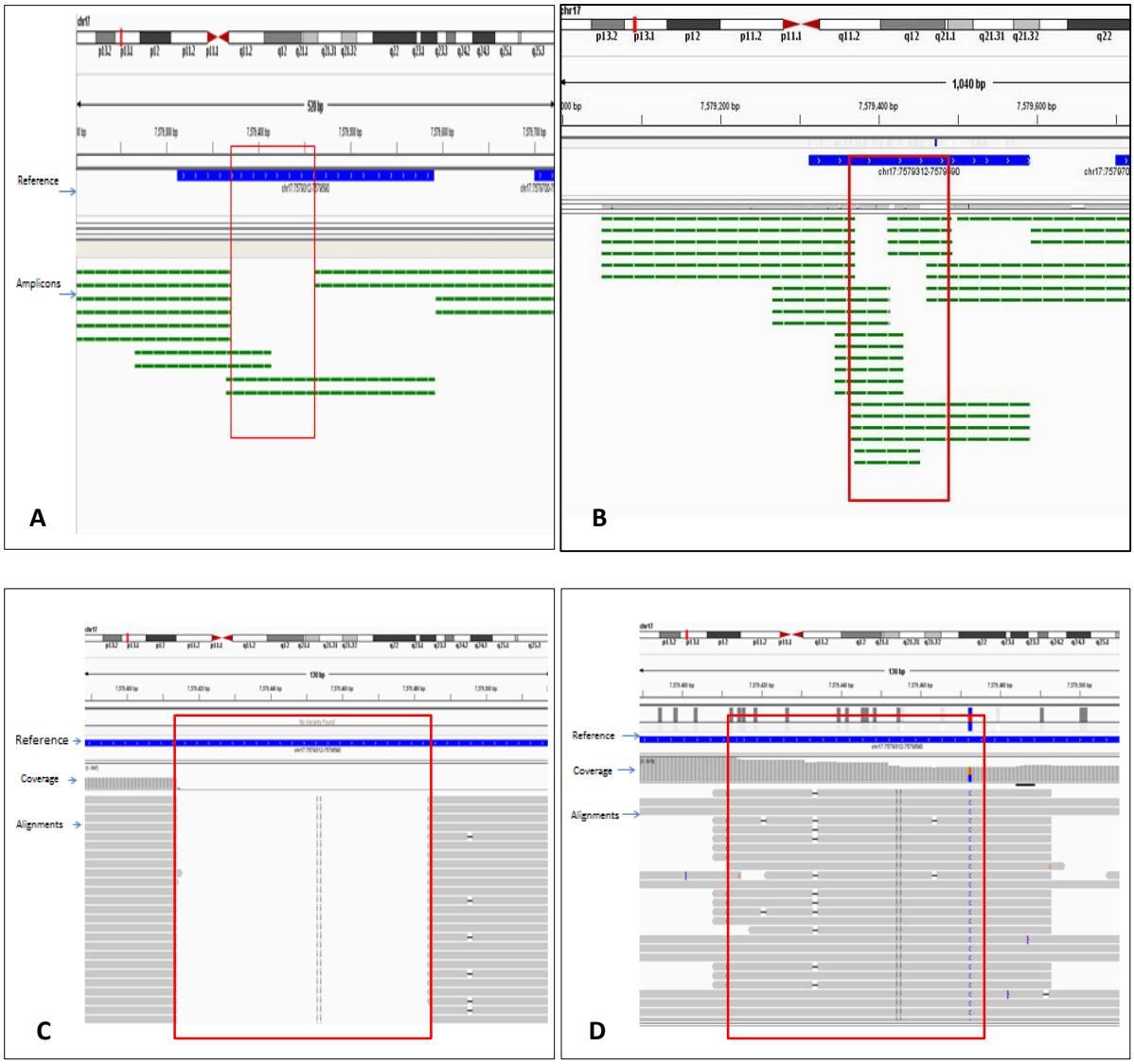


Figure 2.5. **An example of improved target coverage by boosting probe design**

A and **B.** The visualised BED files with the UCSC software show probes (green bars) designed for a targeted exon (blue bar) containing a GC-rich region (in red box) without (**A**) and with (**B**) probe boosting. The IGV files show the alignment of sequence reads (grey bars) across the GC-rich region enriched with the HaloPlex probes without (**C**) and with (**D**) the boosting.

2.3.1.5. Saving cost by reducing the amount of HaloPlex probes used for DNA target enrichment

In this step, HaloPlex probes are hybridised to target DNA fragments in an in solution based hybridisation capture reaction. Unlike the array-based capture which is based on influx of large amounts of DNA and subsequent washing of the off-target DNA on a solid surface to which probes are fixed, the in solution-based capture is using larger amounts of probes on smaller amount of DNA. Through subsequent washing steps, the off-target DNA and the free probes are cleared. As the probes are biotinylated, they tend to be retained in all steps prior to PCR amplification using the magnetic field. After PCR amplification, size selection removes the excess primers and residual probes. This step is crucial before sequencing as ISPs containing only primers or probes consume capacity of the Ion chip and yield unusable short reads. Accordingly, the HaloPlex probes are expensive, they account for about 40% of cost for laboratory consumables of the test.

To cut the cost, we successfully reduced the amount of probes used for the target DNA enrichment, from 20 µl to 17 µl per sample. As shown in Figure 2.6, this reduction (B) produced the same results as the recommended amount of probes did (A) in 8 un-paired CLL-samples.

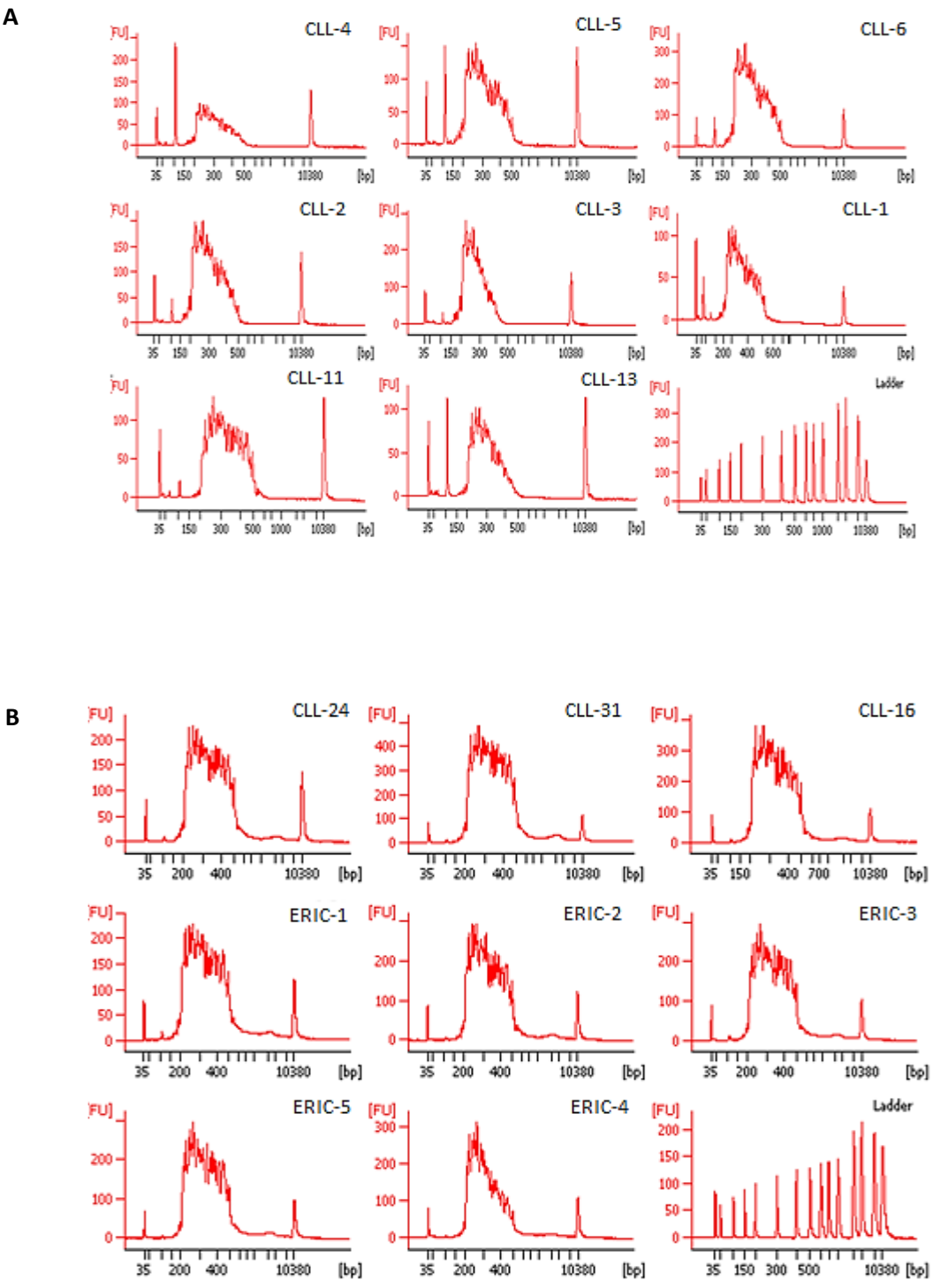


Figure 2.6. **Reduced amount of HaloPlex probes did not affect results of target enrichment**

Bioanalyser electrophoreograms showing size of DNA libraries enriched with recommended 20 µl of HaloPlex probes (**A**) and 17 µl of the same probes (**B**) in 8 unpaired DNA samples.

2.3.1.6. Improvement of efficiency of DNA library size selection

As mentioned in Section 2.2.8.1, the size of digested DNA fragments varies from 100 - 500 bp. With probes ligated the ideal size of the DNA library for sequencing is between 150 - 550 bp. After target enrichment and amplification, it is important to eliminate free probes and primers. Efficiency and quality of the purification is determined by multiple factors, including the duration of incubation of the DNA library with XP-beads and the number of purification procedures. To improve the efficiency, we initially compared the effects of the incubation durations for two CLL libraries. As shown in Figure 2.7. A and B, there were no visible peaks of recovered DNA libraries after incubation with the beads for the recommended 5 minutes at room temperature, while the DNA libraries were obviously recovered with the prolonged incubation for 15 minutes, as presented in Bioanalyser electrophoreograms Figure 2.7.C and D.

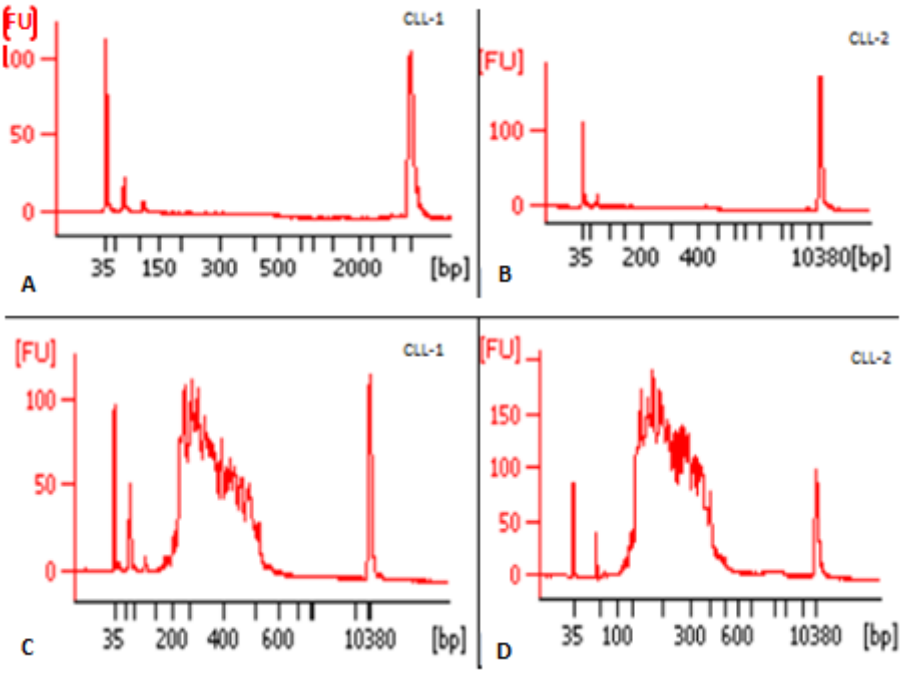


Figure 2.7. **Improved recovery of DNA library after optimised purification step** Bioanalyser electrophoreograms showing no yield with the recommended 5 minutes DNA- XP bead incubation in 2 CLL samples, **A** and **B** efficient recovery was achieved (DNA peaks of 150 - 550 bp) with elongation of incubation to 15 minutes, while some residuals of free primers and probes are left (peaks of ≤ 100 bp), **C** and **D**.

However, these recovered DNA libraries still contained visible free probes and primers (Figure 2.7.C and D). To further remove them, we next repeated the purification procedure. As expected, a satisfied purity of the libraries was achieved with double purification, with no more visible free probes and primers in Bioanalyser electrophoreograms (Figure 2.8.A and B) and an increase in the average size of DNA templates in the library was obtained.

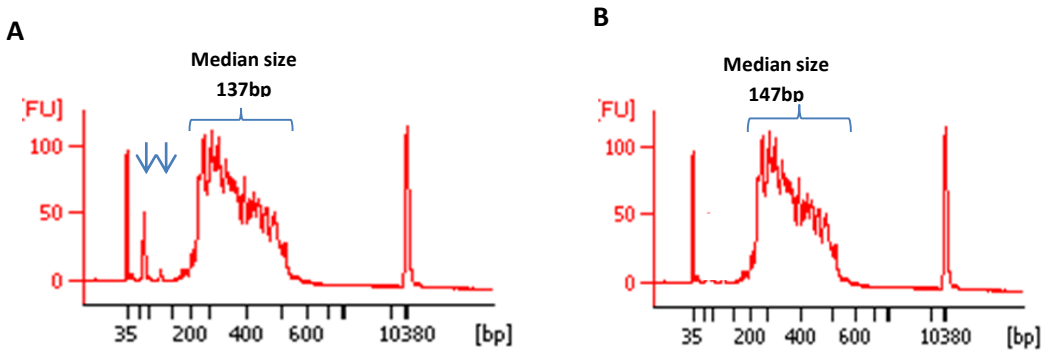


Figure 2.8. **Quality improvement of DNA library with double purification**

Results of purification for one of the two CLL DNA libraries showing the purity achieved with single (**A**) and double (**B**) procedure, as presented in Bioanalyser electrophoreograms. The residual primers and free probes peaks are marked with the two blue arrows. The two narrow fluorescent intensity peaks present two DNA size markers.

2.3.1.7. Optimisation of conducting coverage analysis

It is very important to determine the coverage uniformity and coverage depth (identify regions with lack of sufficient coverage) in a time efficient way to report the sensitivity of test specifically when clinical translation is required. For this purpose, Ion Torrent Coverage Analysis Plugin was used to test different coverage analysis settings including the defaulted, the manufacturer (Agilent technologies) recommended and the custom developed coverage analysis setting. As shown in Table 2.10, the defaulted setting used the whole genome as the target region. The disadvantages of this setting were time ineffectiveness (not feasible for large NGS data) and it created a misleading coverage analysis output (low coverage

compared to whole genome as a target) and regions with lack of sufficient coverage could not be identified without IGV visualisation. The manufacturer's recommended setting required uploading of specific target region BED file created by Sure Design Software. Although, this setting generated a coverage analysis comparative to the specific target region, the output data was in a compressed form in which regions with a lack of coverage could not be easily identified. In contrast, the custom developed setting used an individual target gene strategy to obtain easily analysable information in shorter time periods. In this approach, the individual target gene library BED files were created and uploaded to the server, the tabular information provided for each target gene was not compacted and specific target regions with low coverage could be easily identified and reported without the need to IGV visualisation.

Table 2.10. **Comparisons of performance of various setting of Ion Torrent sequencing coverage analysis**

		Defaulted setting	**Manufacturer's recommended setting**	**Custom developed setting**
Uploaded file(s)	Number	0	1	15
	Type	NA	Whole target BED	Target gene specific BED
Number of coverage analysis run(s) performed/ sample		1	1	15
Output comparative to		Whole genome	Whole target region	Specific target gene
Condition of data		Compressed	Compressed	Not compressed
Identification of regions with low coverage without IGV visualisation		No	No	Yes
Time spent to analyse data (hr.)/ sample		6	6	2

2.3.1.8. Optimisation of stringency of variant calling to improve sensitivity and precision of the test

Besides coverage depth, criteria set for variant calling affect both sensitivity and precision of the NGS test. This is particularly important for testing low level mutations. For this reason we compared analysis results of 5 known mutations in *TP53*, all of which were diluted to 1% VAF and 4 of which to 5% in wild-type DNA, using three different settings: the defaulted high and low stringent settings and a customised setting. Unsurprisingly, the defaulted high stringent setting identified only 3 of the 9 mutations, while the low stringent setting identified all 9 mutations but also 2 false variants. However, with the finely adjusted setting (the different minimum variant Phred quality = 16, minimum mismatched base reads = 10, minimum coverage on either strand = 2, maximum strand bias = 0.95, minimum relative read quality = 8.5 and maximum common signal shift = 0.25), we were able to detect all of the 9 mutations without false negatives and false positives (Table 2.11). An example of the difference in results between the three settings visualised in IGV display is shown in Figure 2.9. The customised setting was the best for both sensitivity and precision, and therefore applied to subsequent studies.

Table 2.11. **Comparison of specificity and sensitivity of various stringency settings of TVC v4.2. for calling 9 *TP53* mutations diluted to 5%-1%**

Settings	Defaulted high stringency	Defaulted low stringency	Customised stringency
Minimum variant Phred quality	10 (P = 0.100)	6 (P = 0.250)	16 (P = 0.025)
Minimum mismatched base reads	20	6	10
Minimum coverage on either strand	3	2	2
Maximum strand bias	0.9	0.95	0.95
Minimum relative read quality	8.5	6.5	8.5
Maximum common signal shift	0.25	0.2	0.25
Maximum homo-polymer length	8	8	8
True positive	3	9	9
False negative	6	0	0
False positive	0	2	0
Results	Sensitivity: 33% Precision: 100%	Sensitivity: 100% Precision: 82%	Sensitivity: 100% Precision: 100%

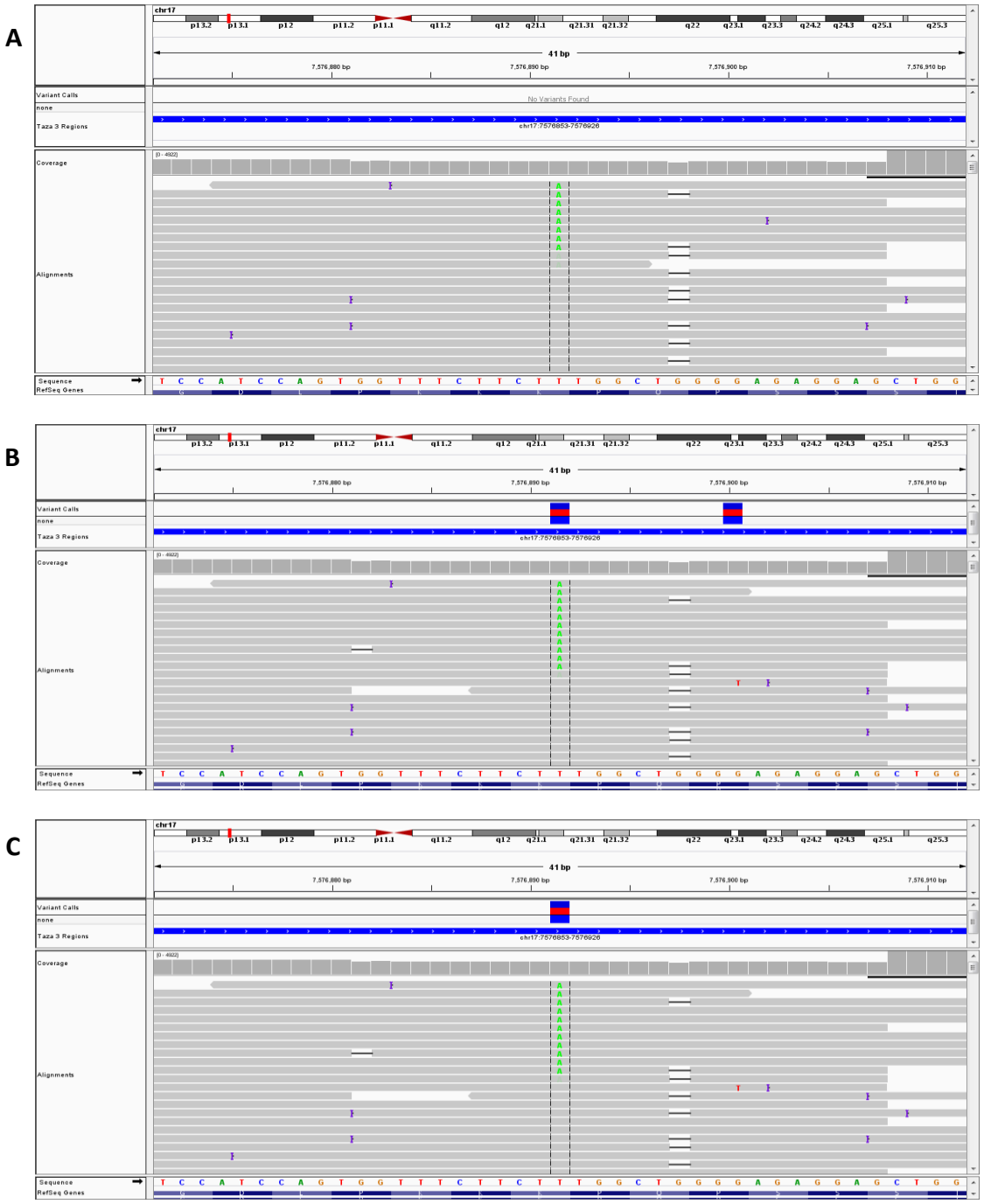


Figure 2.9. **An example (IGV view) of variant detection and variant calling using TVC pipeline for *TP53* mutations using different stringency settings**

The IGV displays show results of a sequencing test for a somatic non-synonymous mismatch in *TP53* (17:17576891T>A, gated in the middle of the displays) of a CLL DNA sample diluted to 1% VAF with wild-type DNA, as analysed with defaulted high stringent (**A**), defaulted low stringent (**B**) and the customised (**C**) settings. Note that the blue-and-red bars above the probe (Taza 3) region bar (blue) are the variant calls, the numbers of which are different for the three settings.

2.3.2. Good test reliability

2.3.2.1. Using known germline SNPs as an indicator

Single nucleotide polymorphisms (SNPs) are germline variants inherited from an individual's parents. They therefore exist in all cells including the tumour cells. There are about 1.42 million SNPs in human genome at a density of 0.47 per 1 Kbp [250]. We therefore, firstly compared this density with that found in our study cohort. With this Ion Torrent PGM sequencing, we detected 21.2 $\pm$ 2.1 (Mean $\pm$ SD) in the target regions (52.324 Kbp) of each sample. The calculated SNP density was indeed, 0.41 $\pm$ 0.04 (Mean $\pm$ SD), which was very similar to the expected.

Because the expected allele frequency of a SNP is 50% in heterozygous state and 100% in homozygous state [251], we secondly used them to exclude test allele bias introduced in PCR amplification, indicated by disparity of their frequencies from 50% or 100%. For this purpose, we summarised distribution of VAF of the total of 233 SNPs identified with the test (Figure 2.10). As expected the VAF was between 40% and 60% and 90% and 100%, although there were a few of them fallen outside of the two peaks which might be due to mosaic chromosomal loss or gain that commonly occur in tumour samples.

Moreover, each individual has a unique SNP pattern which can be used to exclude DNA contamination between samples. We therefore finally compared the pattern in the 11 CLL samples (Table 2.12), which clearly show the difference among this cohort of DNA samples, with no samples sharing a similar pattern. These formed evidence strongly supporting the good reliability of our test.

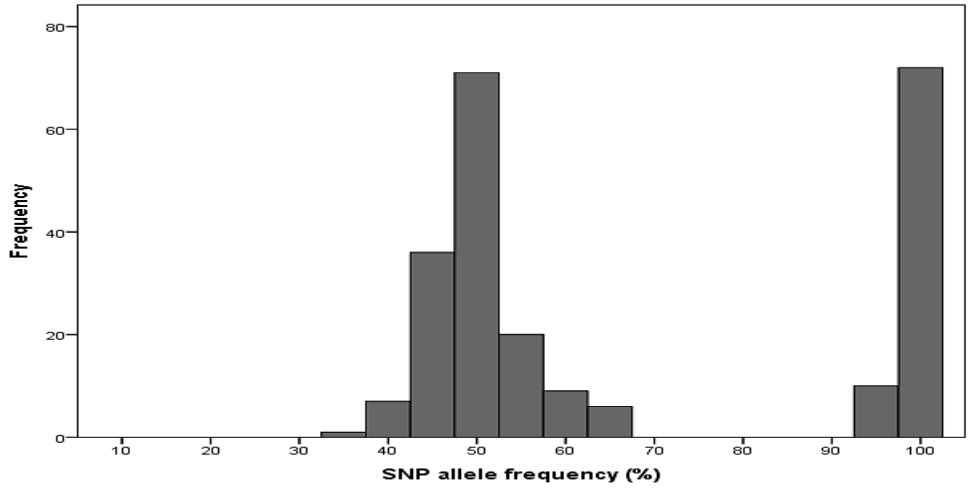


Figure 2.10. **Distribution of identified SNP variant allele frequencies in 12 CLL samples sequenced by Ion Torrent PGM**
A total of 232 SNPs in 11 CLL cases were examined for distribution of allele frequencies. As expected, 97% of the SNPs fallen in between 40 - 60 and 90 - 100 VAF%.

2.3.2.2. The expected transitions/transversions ratio functioned as quality control

Transitions to transversions (Ti/Tv) ratio have been used as quality control in multiple studies [252, 253]. Transitions (Ti) involve interchanges of nucleotide between purines (A and G) or between pyrimidines (C and T), while transversions (Tv) involve interchanges between purines and pyrimidines. Theoretically if these interchanges occur randomly, the chance of Tv changes are 2 folds of that of transitions. But in fact, Ti changes occur more than Tv changes due to their chemical properties. Furthermore, the ratio differs across different regions in human genome. Inside exons, the Ti/Tv ratio is 3 or higher, while it is only 2 inside introns [254]. In target (exon) regions of the 11 samples included in this chapter, this ratio was 3.6 ± 0.4 (Mean ± SD), which was expected and therefore further confirmed the good reliability of our test.

Table 2.12. **SNP patterns in 11 different CLL DNA samples**

Variant info.	R	V	dbSNP	CLL-1	CLL-4	CLL-6	CLL-11	CLL-12	CLL-15	CLL-17	CLL-18	CLL-22	CLL-24	CLL-32
2:61719275	T	C	rs3816341					52						
2:61722724	G	T	rs6171632					100						
2:141032088	C	T	rs1386356	49	100	100	50	50	100	48	100	48	51	100
2:141072519	C	G	rs17386226											54
2:141092084	T	G	rs79879036								53			
2:141116420	C	T	rs150879175								48			
2:141116447	G	T	rs35546150				47					45		
2:141128779	C	G	rs76554185						52					
2:141130695	C	T	rs16843864		47				44		46			
2:141232800	C	T	rs72899872										45	48
2:141242918	T	C	rs34488772						52					
2:141245204	T	C	rs74789055						46					
2:141259283	G	A	rs35296183		40				47	45			99	99
2:141259376	G	A	rs35164907		48				100	46			100	100
2:141260668	A	G	rs4444457	50	49	53	50	50	100	100		51	100	100
2:141267573	G	A	rs35821928	60			64					70		
2:141272253	C	T	rs61732738								47			
2:141274504	G	A	rs75124368											100
2:141274576	T	C	rs4954672	50	99	47	51	99	100	47	97	49	98	97
2:141457962	C	T	rs34694228			52								
2:141457985	T	A	rs13431727		41								97	99
2:198265526	A	G	rs788018	100	100		46	49		100	53	48	54	
2:198274685	C	T	rs377192403			43								
6:26157073	A	G	rs2298090							47				
7:82435033	C	T	rs12668093	54	41						49		49	51
7:82451836	G	A	rs146099474		41									
7:82453708	A	C	rs2522833	100	100	53	51		50		44	51	51	48
7:82544987	A	G	rs17156844		57	54	51	51			99	49		
7:82579183	T	C	rs201013392										44	
7:82580293	A	T	rs199626449											
7:82581859	C	T	rs976714	57	58	49	51	51	54	46		48		
7:82582258	T	G	rs10261848											
7:82582846	C	T	rs10954696	49	59	50	48	46	50	48		49		
7:82583280	C	T	rs17148149											
7:82583388	A	G	rs10487647										100	
7:82583609	A	C	rs10487648	37				47	44					
7:82584574	G	T	rs61995911				42					48		
7:82585803	G	C	rs114445550											
7:82595324	C	T	rs9969358										46	
7:82595742	T	C	rs28680905								45			
7:82764425	C	G	rs2877	58	100	100			100	55	100		45	
7:82784456	A	G	rs6972461	47	38		100	100		44	45	100		44
7:82784501	G	T	rs201808333											
7:82785097	T	C	rs61741659						46	45	46		48	44
7:82785099	T	G	rs61741653				48	50				52		
7:124481185	C	A	rs 35536751				48						50	
8:106813518	C	G	rs11993776	49			100		50			100		50
8:106813672	A	G	rs920628				66					73		
8:106813888	G	T	rs200643137											
8:106814086	T	C	RS16873732				64					64		
8:106814279	A	G	rs28374544											
8:106814656	G	C	rs2920048					52		53				
8:106814695	C	G	rs355998713				67					65		
8:106815286	T	C	rs1442320				62					61		
8:106815474	C	T	rs16873741							100				
8:106815517	C	T	rs11995760			100								
8:106815679	A	G	rs16873744			48								
9:139390958	T	C	rs11574911				45							
9:139391636	G	A	rs2229974		**50**	100	50	53	100	52				100
11:102207851	G	A	rs1055088	100			46	99	100		100	50	100	100
11:108160350	C	T	rs1800058			57								
11:108163487	C	T	rs1800889	44										
11:108175462	G	A	rs1801516			50								
11:108183167	A	G	rs659243	100	100	100	100	100	100	100	100	100	100	100
15:93510603	A	G	rs4777755	99	99	48	98	98	99	53	99	98	53	49
15:93521604	A	G	rs11074121	100	99	35	100	99	99	53	99	98	43	47
15:93536197	C	T	rs2272457	47	100			47	55					49
15:93552488	C	T	rs34315566										51	
17:7579472	G	C	rs1042522	**97**	**97**	96	64	49	61	58	97			97

2.3.3. Sensitivity of the test

2.3.3.1. Serial dilution test

Analytical sensitivity designates both the lowest limit of detection and the lowest limit of quantitation. The former is likelihood of a test to reliably distinguish the lowest analyte concentration from blank. While the latter not only include reliable detection, but also the lowest concentration measurement without bias or imprecision [255]. Both of them were important for successful development of our NGS test. Balanced between clinical requirement and test cost, we set up the limit of detection of this test at 1% VAF as mentioned in Section 2.3.1.3. We then tested this limit in a serial dilution experiment, in which 4 known mutations in *TP53* were diluted down to 0.2% VAF with wild-type DNA.

We selected 4 CLL samples that carried 4 distinct *TP53* point mutations (55% - 97% VAF) as detected by FASAY and Sanger sequencing. One of the 4 samples contained an extra mutation at only 5% as found with the deep NGS. As mentioned above, each of these 5 point mutations was diluted to 20%, 5%, 1 % and/or 0.2%. As shown in Table 2.13, all of these mutations diluted to 20%, 5% and 1% were detected under the optimised test conditions. Even at the lowest level of 0.2%, the mutations were detected in 4 out of the 5 samples. Regarding the accuracy, the actual levels measured with the NGS method were very close to the expected, being 20.33 ± 5.78, 4.27 ± 0.82, 1.16 ± 0.14 and 0.15 ± 0.06 (Mean $\pm$ SE), respectively, as shown in Figure 2.11. These results provided strong evidence that our test is able to detect mutations down to at least 1% with a high accuracy.

Table 2.13. **Details of the point mutations tested in a sensitivity test experiment using HaloPlex and Ion Torrent PGM techniques***

ID	*TP53 point* mutations in the original samples using FASAY and Sanger sequencing						Obtained data for serial dilutions to											
							S-20 %			S-5 %			S-1 %			S-0.2%		
	Exon	Chr. location	Ref	Var	VAF %	Codon	VAF	VQ	Cov.	VAF	VQ	Cov.	VAF	VQ	Cov.	VAF	VQ	Cov.
CLL-6	5	17:7578394	T	C	98	179 H>R	21%	2451	1987	4.90%	96	2097	1.0%	26	3434	0.30%	26	1565
CLL-A	7	17:7577539	G	A	96	248 R>W	10%	668	2244	2.20%	33	4365	1.0%	32	3211	0.04%	19	4261
CLL-4	8	17:7577100	T	C	74	280 R>G	30%	1373	600	3.90%	20	636	1.7%	33	1004	0.00%	….	711
CLL-1	9	17:7576891	T	A	24	319 K/X	NA	……	…..	6.10%	127	1370	1.0%	30	2273	0.06%	29	1531
	9	17:7577079**	C	A	5	287 E/X	NA	……	….	NA	….	…..	1.1%	30	638	0.19%	26	1013

* For each variant allele (Var.), the reference allele (Ref.), the chromosomal location (Chr. location) using Human reference sequence hg19, the exon and the codon change for each CLL sample using the identification (ID) code are shown. The % of the obtained variant allele frequency (VAF) as well as the variant call Phred Quality score (VQ) and coverage depth (Cov.) for each particular nucleotide are presented. CLL-6, CLL-A and CLL-4 had mutations detected by FASAY and Sanger sequencing. CLL-1 (harboured two distinct point mutations) was not included (NA) in the 20% serial dilution sample. Moreover, in the 5% serial dilution sample (S-5%), the mutation with original 5% allele frequency is not expected to be found at 5% level and hence its allele frequency lags by 5 folds behind the dominant mutation which was located at codon 319. CLL -24 which had Wt. TP53 gene was used as the diluent.

* * This mutation was detected by re-screening of the sample by Ion Torrent PGM

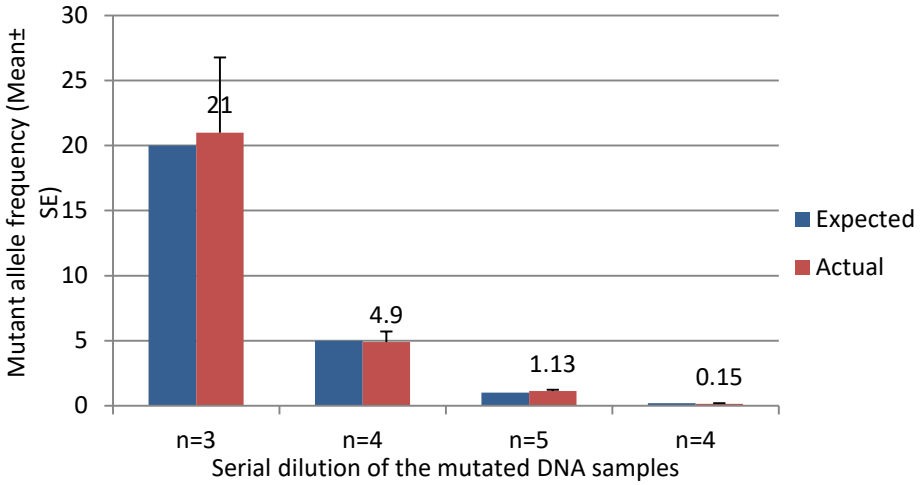
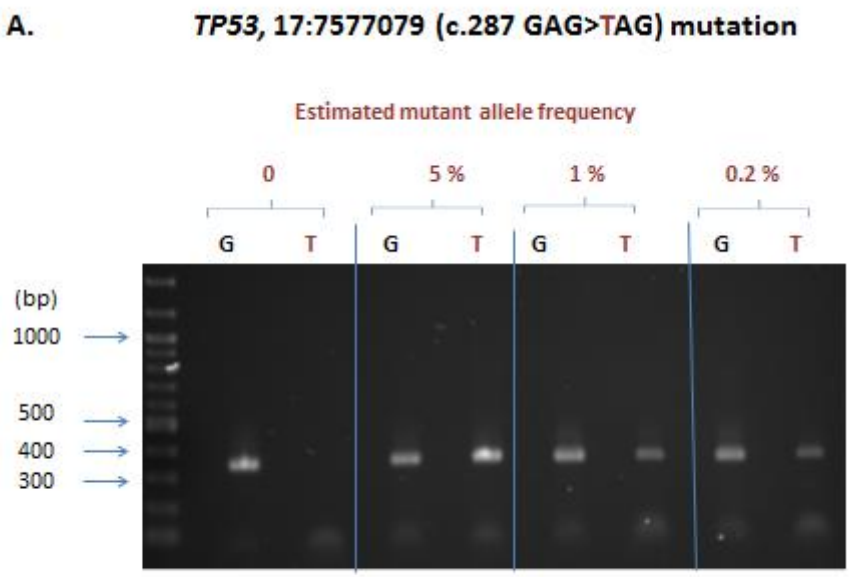


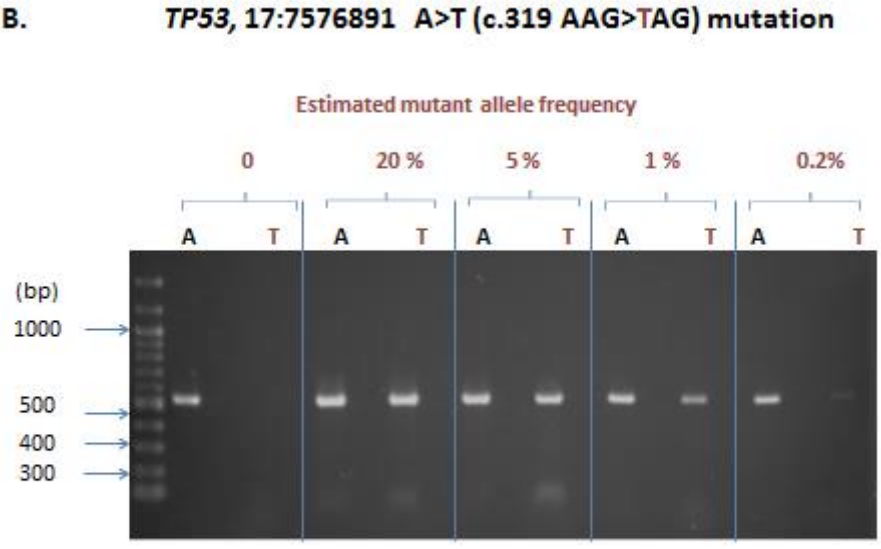
Figure 2.11. **Serial dilution of *TP53* point mutations**

The 4 artificial samples were created by mixing and serially diluting 4 CLL samples with 5 distinct *TP53* point mutations in a *TP53* wild type DNA to obtain allele frequency of 20%, 5%, 1% and 0.2% for each variant in these samples, respectively. The optimised stringency settings of TVC variant caller (v4.2) facilitated the readily mutation identification without increase in false positive calls.

2.3.3.2. Confirmation of low-level mutations identified by the NGS using AS-PCR

As mentioned in Section 1.3.2.4, AS-PCR is highly sensitive for confirmation or detection of known mutations [256]. We employed it to confirm mutations at a low level around 1% detected by NGS. In order to avoid both false positive and false negative results, the PCR conditions had been carefully optimised before use by adjusting the amount of DNA, concentrations of primers, $MgCl_2$ and dNTP with or without addition of DMSO, annealing temperature and/or cycle number of PCR, as shown in Table 2.6. Using the optimised AS-PCR, we successfully confirmed all of the 4 *TP53* mutations in the patient samples still available at a level of 1% and 0.2% VAF as diluted in wild-type DNA, as shown in Figure 2.12. Notably, AS-PCR product amplified from 0.2% of a mutation, which was undetectable at that level with the NGS method, was also visible, although not as strong as those from the same level of other mutations (Figure 2.12.D). These results steadily confirmed the fidelity of mutations detected with the NGS method at the level of detection limit (1%).


A.
TP53, 17:7577079 (c.287 GAG>TAG) mutation
Estimated mutant allele frequency
0
5 %
1 %
0.2 %
G
T
G
T
G
T
G
T
(bp)
1000
500
400
300


B.
TP53, 17:7576891 A>T (c.319 AAG>TAG) mutation
Estimated mutant allele frequency
0
20 %
5 %
1 %
0.2%
A
T
A
T
A
T
A
T
A
T
(bp)
1000
500
400
300

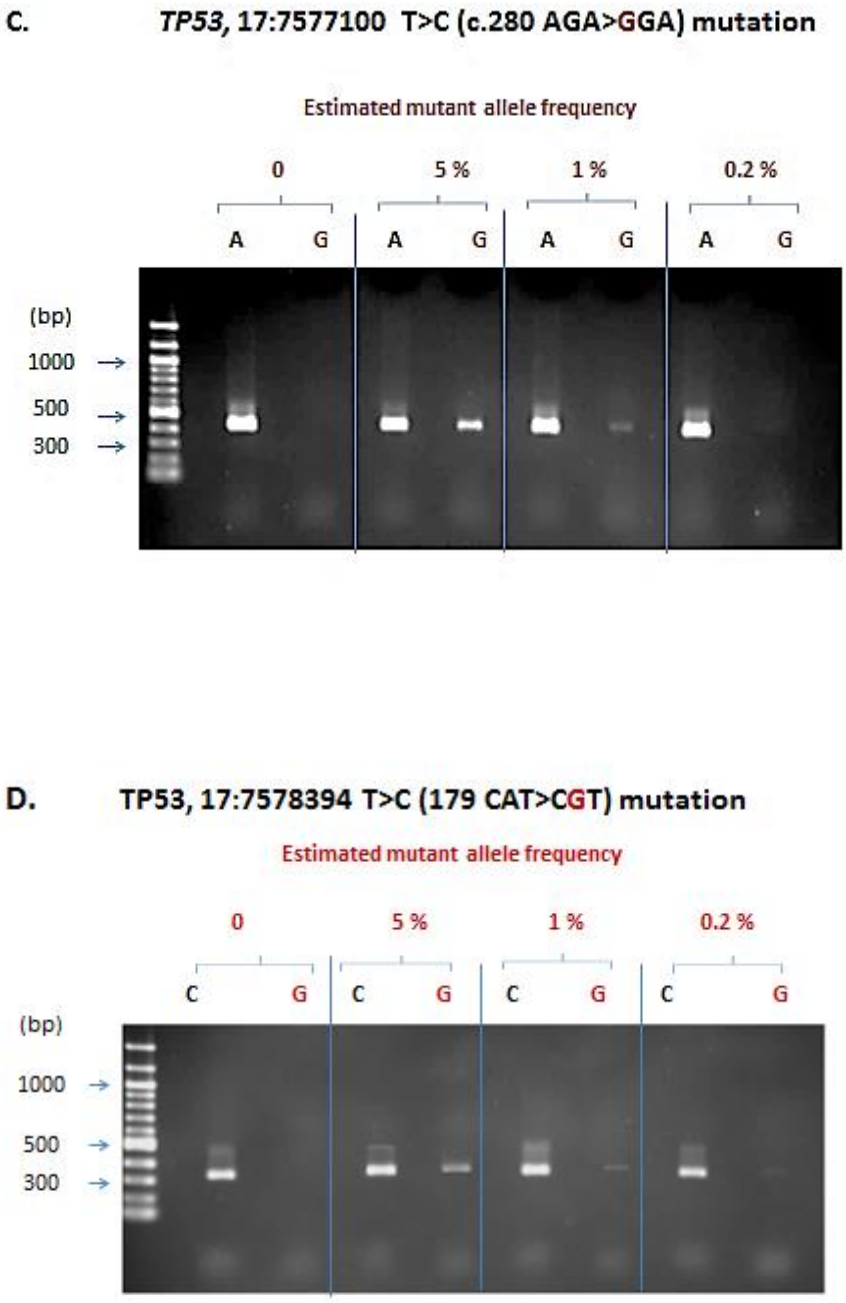


Figure 2.12. **Agarose gel electrophoresis images for validation of NGS-detected *TP53* point mutations by AS-PCR**

As described in (Section 2.3.3.2.), serially diluted mutated DNA was used for the AS-PCR. **A.** PCR bands of 338 bp were amplified from wild-type (G) and mutant (T) alleles. **B.** PCR bands of 527 bp from wild-type (A) and mutant (T) alleles. **C.** PCR bands of 427 bp from wild-type (A) and mutant (G) alleles, and **D.** PCR bands of 334 bp from wild-type (A) and mutant (G) alleles.

2.3.3.3. High repeatability of the Ion Torrent PGM sequencing test

Although most conditions for the NGS test were controllable, the sequencing on the PGM might vary due to conditions out of our own control, e.g. template DNA amplification on beads and loading on chip. It was therefore necessary to test repeatability of the sequencing results. To do so, we sequenced DNA libraries prepared from 4 patients under the optimised conditions on separate Ion chips in different runs. The total number of variants called, including both germline and somatic variants, in target regions was very similar between the two runs (Table 2.14). Moreover, statistical analysis revealed a good linear correlation of the mutation VAF% between the two runs for all of the 4 DNA libraries as shown in Table 2.14 and Figure 2.13. Notably, such a correlation existed even for mutations with VAF lower than 20% (see those in red boxes in Figure 2.13).

Table 2.14. **Summary of Ion Torrent PGM reproducibility test performed for 4 CLL samples in 2 sequencing runs**

CLL-ID	Runs	Number of variants			R^2	Regression	P
		Germline and somatic synonymous	Somatic non-synonymous	Total			
CLL-1	A	22	8	30	0.977	B=2.89+0.95A	<0.001
	B	22	7	29			
CLL-4	A	29	3	32	0.966	B=1.18+1A	<0.001
	B	29	3	32			
CLL-6	A	23	2	25	0.991	B=2.08+1.02A	<0.001
	B	23	2	25			
CLL-24	A	24	1	25	0.996	B=1.25+0.96A	<0.001
	B	24	1	25			

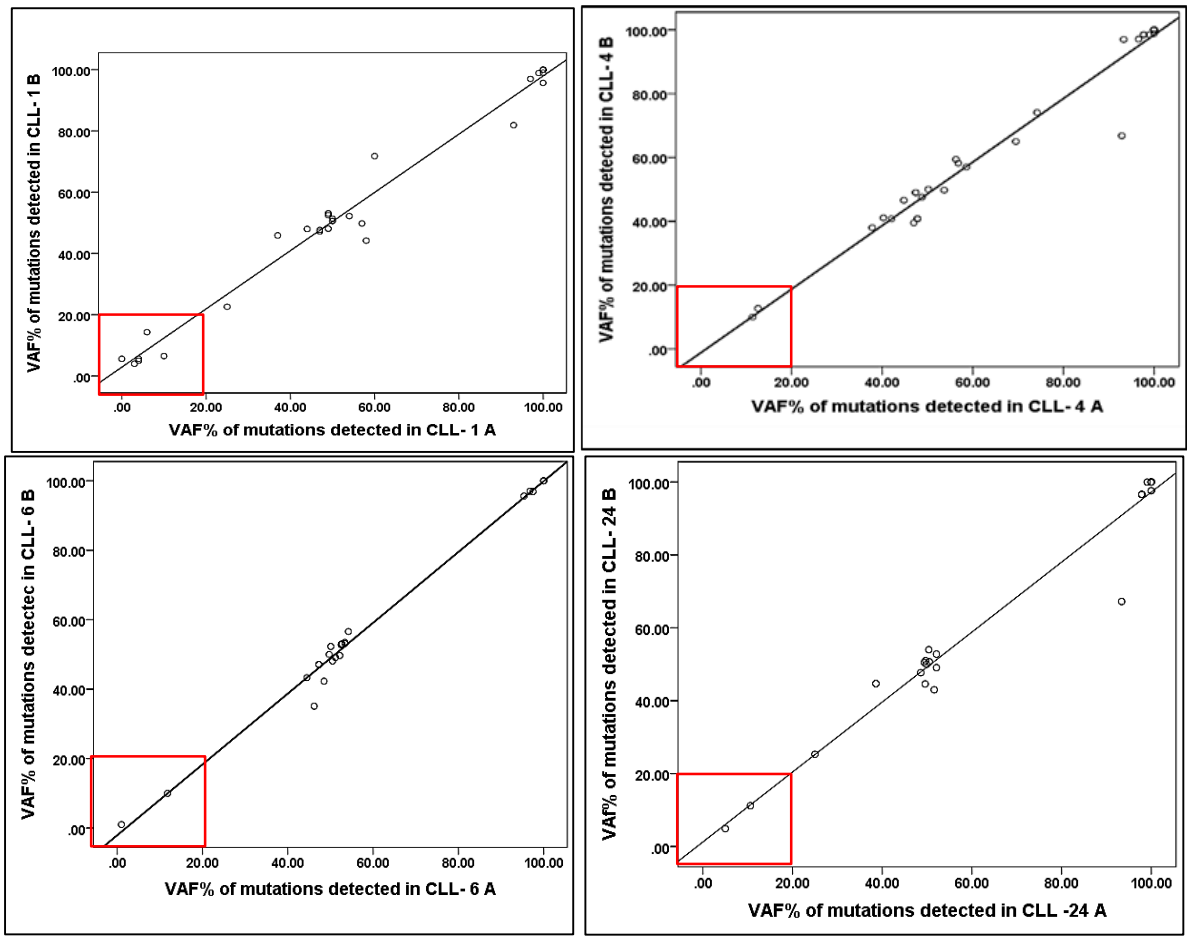


Figure 2.13. **Comparisons between VAF% of all the mutations detected in replicates of Ion Torrent PGM experiments in 4 CLL samples**
Statistical analysis showing a positive linear correlation between the frequencies of mutant alleles detected in repeated experiments for 4 CLL samples (CLL-1, CLL-4, CLL-6 and CLL-24) using Ion Torrent PGM method.

2.3.3.4. Good reproducibility confirmed with alternative methods in multi-centre blind studies

Determining inter-operator variability is an essential step for establishing a new method as it is an approach to confirm the reported results with alternative technologies. Moreover, double blind multicentre studies reduces risk of bias for all types of variants, including indels, which are more likely to be false positive as detected with the Ion Torrent PGM technique [257, 258]. For that reason, we sent 7 DNA samples of our study cohort to the group led by

Professor Stankovic in University of Birmingham to screen mutations in *ATM* with alternative methods. Among these samples four had a total of 6 mutations (all > 30% VAF, including a deletion of 10 nucleotides), one had 2 mutations including a point mutation with 14.7% VAF and a single nucleotide insertion with 14.4% VAF, and two had no mutations in target regions of this gene as tested with our NGS method (Table 2.15). Without any information provided, the Professor Stankovic's laboratory confirmed all of the 6 mutations with VAF > 30% with DHPLC and Sanger sequencing, with no mutations detected in the two wild-type samples. Not surprisingly, the 2 mutations with < 15% VAF were not confirmed, most likely due to the detection limit of Sanger sequencing.

In a double blinded multicentre study organised by the European Research Initiative on CLL (ERIC), 5 anonymous CLL samples were sent to us and 45 other centres over the world to screen for mutations in *TP53* using different methods, including DHPLC + Sanger sequencing, FASAY + Sanger sequencing and NGS methods. Using the Ion Torrent PGM method, we detected all of the 5 mutations in 4 samples, as identified by ERIC using DHPLC and Sanger sequencing. In addition, we detected an extra mutation with the VAF being only 4.3% in sample 4 (Table 2.16). More importantly, the VAF of mutant alleles measured with our NGS method correlated very well with that recorded by ERIC, with the correlation coefficient being 0.99, $P < 0.001$, slop in regression equation close to 1 (0.972) and the intercept as small as 3 (Figure 2.14). Those results further confirmed that results generated with our NGS test are well reproducible by conventional methods, although low-level mutations were not included due to the limited sensitivity of those alternative methods.

Table 2.15. **Summary of *ATM* mutations detected by the University of Birmingham and the Ion Torrent PGM technique** *

CLL-ID	*ATM* sequence information detected by Ion Torrent PGM										Confirmed by DHPLC and Sanger
	Mutational status	Exon	Chr. location (hg 19)	Ref.	Var.	c. DNA	Codon	VAF %	VQ	Coverage	
CLL-11	Mu	10	11: 108121763	G	A	c.1571	p.524 W/X	37	5033	1582	Yes
		57	11:108213973	G	A	c.8293	p.2765 G>S	38	5478	1655	Yes
CLL-12	Mu	41	11:108186598	T	C	c.6055	p.2019 Y>H	98	8523	559	Yes
CLL-15	Mu	22	11:108143528	T	G	c.3233	p.1078 L>R	14.7	1898	2906	No
		33	11:108168106-07	-	C	c.5001-2	p.1668 L/fs	14.4	126	529	No
CLL-17	Mu	41	11:108186599	A	G	c.6056	p.2019 Y>H	47	9822	2354	Yes
		58	11:108216582-86	TTTGG	---	c.8531-35	p.2845 W/fs	48	4797	1483	Yes
CLL-18	Mu	48	11:108198445-54	TAGAAAATCC	---	c.7049-58	p.2351 E/fs	31	2185	1553	Yes
CLL-22	Wt										Yes
CLL-32	Wt										Yes

*Details of chromosomal location (Chr. location) using human reference (hg19). Reference allele (R.), variant allele (V.), c. DNA position, protein code and exon harboured the mutations already detected by other centres using DHPLC and Sanger sequencing in ATM gene for 4 CLL samples (A-D) are shown. The variant allele frequency (VAF %), variant Phred quality score (VQ) and coverage depth are illustrated for each variant detected by Ion Torrent PGM. As only the protein coding sequences of the gene are targeted in the PGM method, the sequence information for the two mutations located in exon 1 (non-coding exon) could not be obtained.

Table 2.16. ***TP53* gene mutations detected by ULM, Germany and the Ion Torrent PGM technique (our lab)***

CLL ID	Ion Torrent PGM (our laboratory)										DHPLC, Sanger (ERIC)	
	TP53 mutational status	Exon	Variant chr. locarion hg 19	R.	V.	c.DNA	Codon	VAF %	VQ	Coverage	*TP53* status	VAF %
ERIC-1	Wt										Wt	
ERIC-2	Mu	7	17:7577566	T	C	c.715	p.N239D	54.2	11531	1999	Mu	44
ERIC-3	Mu	5	17:7578413	C	A	c.517	P.V173L	41.4	4220	1125	Mu	47
ERIC-4	Mu	6	17:7578203	C	G	c.646	p.V216L	4.3	65	2000	Mu	------
		8	17:7577127	C	A	c.811	p.E271X	45.6	4548	1052		41
		4	17:7579317	A	G	c.370	p.C124R	39.9	5880	1660		38
ERIC-5	Mu	10	17:7573999	T	-	c.1028	p.E343fsX	91.6	19291	1435	Mu	93

*Details of chromosomal location (Chr. location) using human reference (hg19), reference allele (R.), variant allele (V.), c. DNA position, protein code and exon harboured the mutations detected by other centres using DHPLC, Sanger and NGS technique in *TP53* gene for 5 CLL samples (ERIC1- ERIC5). The variant allele frequency (VAF %), variant Phred quality score (VQ) and coverage depth are illustrated for each variant detected by Ion Torrent PGM.

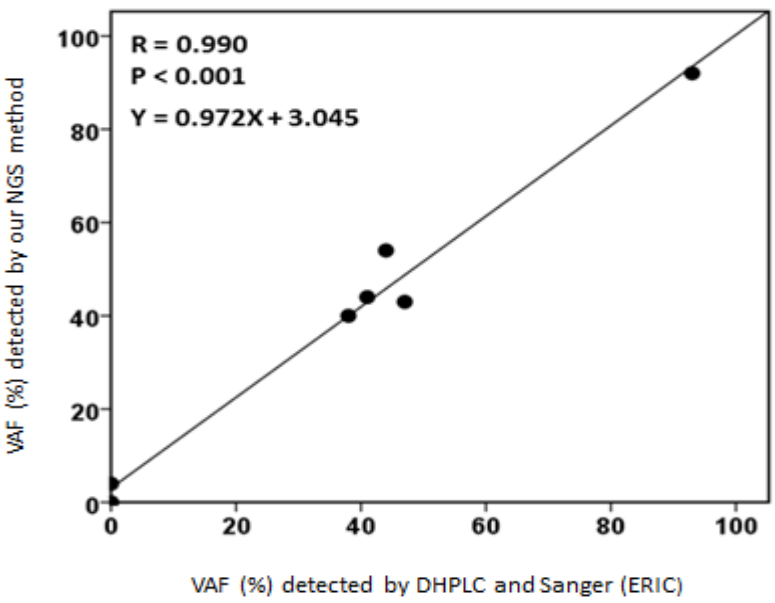


Figure 2.14. **Comparison of mutant *TP53* VAF% detected with our Ion Torrent PGM methods to the blinded VAF% identified with DHPLC and Sanger sequencing by ERIC**

Statistical analysis showing a positive linear correlation between the frequencies of mutant alleles detected with the standard ERIC method and our Ion Torrent PGM method.

2.4. Discussion and conclusion

In this chapter, we aimed to develop a highly sensitive and clinically useful next generation sequencing approach using Ion Torrent PGM to screen recurrent mutations in a panel of genes in CLL. The high sensitivity is required to detect mutations with low variant allele frequencies which are undetectable with conventional methods or even with high throughput NGS techniques with limited coverage depth [259]. The detection of low allele frequency variants is particularly useful to study CLL mutant clonal evolution and for molecular monitoring of patients at early stages and throughout the disease course. Not only in CLL, but also in various types of haematological malignancies and solid tumours, there is still a clinical need for accurate, reproducible methods to measure treatment

response and adequately monitor tumour burden during treatment in order to identify patients who might benefit from individualised therapy [260].

By examining the feasibility of this NGS technique in our laboratory and optimisations of procedures step by step in this chapter, we showed that this method is practically applicable to studies in next chapters which include monitoring clonal evolution in CLL.

The gene panel selection and particularly targeting the disease related coding sequences, provided reasonable coverage. This was important as valuable genetic information is enriched within the coding sequences of genes [261]. The optimised HaloPlex target enrichment design provided uniform coverage across a problematic GC rich target region in *TP53* as it increased the number of probes and amplicons overlaid this region. This problem has been commonly reported in other studies of next generation sequencing [262]. For this reason, efforts have been made to modify PCR conditions [263, 264], however no significant improvement in the uniformity and coverage were obtained. This might be due to that different PCR conditions required for different template DNA amplification.

Reduction in the amount of probe being used for enrichment of each sample, not only was useful for better purification, it also reduced the total cost of enrichment per sample. The incorporation of double purification step with AMPure XP beads (yet with the same cost by the strategy of halving the amount of DNA library purified) resulted in improvement of the read length and total elimination of probes and preserved the chip capacity for deeper coverage of useful bases. Engaging to this approach facilitated saving of up to 16% of the probes which was sufficient for enrichment of 8 extra samples.

Another problem in high throughput short read sequencing is the occurrence of false positive or false negative errors in identifying variants. For this reason, increasing the depth of coverage has been proposed so as to reduce regions with low coverage supply and false negatives. However, coverage depth is not the sole quality filter for variant detection. Various sequencing and mapping criteria can affect the identification of variants specifically for somatic mutations with low allele frequency [265-267].

This process tends to be complicated because of many facts: firstly, tumour samples are enriched in chromosomal abnormalities such as loss of heterozygosity and chromosomal

rearrangements; besides, they tend to be mixed with normal cells by which the proportion of tumour cell alleles to the total number of alleles is attenuated. Additionally, it is not always possible to isolate specific tumour cells before sequencing especially when manipulating clinical samples, as this procedure is time consuming, adds extra cost and does not guarantee 100% purification. Moreover, low level mutations are difficult to be distinguished from background error noise. Therefore, testing the stringency of variant calling and pinpointing the extent to which each filter includes and excludes true somatic mutations with low allele frequency in tumour clinical sample was essential.

Based on optimisation of stringency of variant calling (TVC plugin software by adjusting the parameter's thresholds) for various known true variants with low allele frequency (range 20-0.2%), we could increase the test sensitivity and precision in mutation detection. This enabled accurate detection of mutations diluted to allele frequency as low as 1% with high precision which was further validated by highly sensitive AS-PCR.

We also demonstrated the high reproducibility of the technique by testing different samples starting from DNA preparation step in the same sequencing experiment or from replicates of sequencing experiments. The high concordance of the PGM results with multicentre results in blinded studies for various genes and various types of mutations (including point mutations and short deletions and insertions) imply the reliability and robustness of the method.

In conclusion, we established a highly sensitive NGS test by optimising conditions for both the target DNA enrichment with HaloPlex and analysis of Ion Torrent PGM sequencing data. This was necessary for us to reach the subsequent study goals set up for subsequent chapters of this thesis to establish mutation profile and to study clonal evolution of mutant CLL clones in patients with advanced CLL.

Chapter 3. Targeted gene mutation profile and chromosomal copy number changes of patients with progressive and/or chemo-resistant CLL using ultra-deep NGS and array based whole genome profiling

3.1. Introduction and aims

As mentioned earlier in Chapter 1, CLL is characterised by a heterogeneous disease course in terms of disease progression and response to therapy. It has been suggested that somatic genomic aberrations contribute to this clinical diversity.

Shortly before the start of this work, two independent studies using whole genome and whole exome sequencing approaches had found a number of recurrently mutated genes irrespective of clinical stages of CLL [148, 149]. In addition to those in the two previously well studied genes *TP53* and *ATM,* the two studies had identified novel mutations in other genes including *SF3B1, NOTCH1, MYD88* and *BIRC3*. Similar to *TP53* and *ATM* mutations, a possible link of these novel gene mutations to biology as well as diverse clinical course of the disease was proposed [150, 245]. Moreover, identification of the main chromosomal copy number changes including del11q, del13q, del17p and trisomy 12 by FISH has been used to prognosticate CLL patients. A strong association between del17p and *TP53* mutations has been found. Recent studies have combined FISH and more recently low resolution array based hybridisation techniques with gene mutation analysis to refine the prognostic stratification. This has led to identification of novel chromosomal copy number abnormalities which have been linked to poor prognosis. Largely because of the low sensitivity nature of the sequencing and the array techniques employed, information on the actual incidence, co-occurrences of these lesions, clonal or subclonal architecture of somatic mutations and copy number changes of progressive and or chemo-resistant CLL was scarce.

In parallel with our study, in 2013, a newer study employing whole exome sequencing revealed that heterogeneity of gene mutation architecture and chromosomal copy number changes at early stages possibly fuel CLL clonal evolution over time of disease progression and relapse [198]. Thereby, we hypothesised that the genetic alterations that exist at later stages of disease are expanded from earlier stages and they contributed to disease progression and/or therapy resistance. These alterations are therefore likely to be targets

requiring monitoring from an early stage of the disease. The aims of this chapter were primarily to apply the fast, affordable and sensitive next generation sequencing technique developed based on HaloPlex target enrichment system and Ion Torrent PGM to detect and build up mutation profiles in the panel of 15 genes at a late stage of the selected CLL cases. The secondary aim was to study whole genome copy number alterations (CNA) using high resolution CytoSNP-850K Illumina Bead Chip array to identify possible associations between gene mutations, chromosomal copy number changes and clinical outcome.

3.2. Materials and methods

3.2.1. CLL samples and criteria for case selection

All of the cases included in this chapter were from the Liverpool CLL Bio-bank. They were consented patients and anonymised for research. They were selected to meet the following criteria: firstly, they were diagnosed as typical CLL and had a history of progressive and/ or chemotherapy resistant disease according to the updated IWCLL criteria published in 2008 (as described in Sections 1.1.4 for typical CLL diagnosis and 1.1.6.2 for disease progression and treatment resistance); and secondly, they had serial samples taken not only after disease progression and/or therapy resistance, but also at earlier stages, so that the study on clonal evolution as planned for Chapter 4 could be performed.

There were 120 cases in the Bio-bank meeting these criteria. Among them 32 were selected because they fulfil regulations of the Bio-bank that only release samples with more than 10 vials in storage at any time points of sampling. The cryopreserved CLL cell samples taken at the latest time from these 32 patients were used for study in this chapter. As shown in Table 3.1, an additional sample from a stable (indolent) CLL (CLL-33) was used as a control. In this cohort, the median age at diagnosis was 62 (range: 33 - 82), while median age at sampling was 67.5 (range: 35 - 86). Each of the 32 patients had a history of progressive disease that required initiation of therapy. 22 of them had received therapy at time of sampling, while 10 of them had not received the treatment when sampled. According to IWCLL (2008), 17 patients showed resistance to treatment (refractory to chemotherapy). Only 6 patients partially responded to treatment but for short periods (6 – 24 months).

Table 3.1. **Clinical and laboratory features of the CLL cases used in this study**

CLL cases^	Gender*	Age at diagnosis	At sampling								First treatment***		
			Age	Stage **	No. of lymphoid regions involved	Hb (g/dl)	Lymphocytes 10^3/µl	Platelets 10^3/µl	*IGVH* % dissimilarity	LDT in months	Received	Regimen x cycles	Response/ duration
CLL-1	M	50	63	C	5	9.8	152	99	2.04	<5	Y	Flux6	NR
CLL-2	F	59	70	B	3	10.8	19.4	236	0	<12	Y	Clbx2	NR
CLL-3	F	51	63	B	3	12.5	104.8	221	7.5	<5	Y	FCRx4	PR/9m
CLL-4	M	54	55	B	3	12.3	105.2	201	0	<5	Y	Clb,Rx6	-
CLL-5	M	51	55	B	4	14.2	166	156	0.34	<4	Y	-	-
CLL-6	M	62	67	B	3	11.5	61.9	110	3.47	<3	Y	Clb x6	NR
CLL-7	M	74	79	B	3	11.5	57	315	0	<4	Y	Clbx6	NR
CLL-8	M	60	65	C	1	14.5	19.7	59	0.35	<3	Y	CHOPx6	NR
CLL-9	M	62	69	B	4	12	126.3	211	3.72	<12	N	NA	NA
CLL-10	F	66	68	C	-	9	51.5	301	-	<4	N	NA	NA
CLL-11	F	73	77	B	3	10.4	198.3	183	1.39	<8	Y	Clb,FCx4	NR
CLL-12	M	57	62	C	2	11	800	81	0	<12	Y	Flu x6	NR
CLL-13	M	57	67	C	-	12.1	132.8	46	0.34	<6	Y	Clbx6	PR/7m
CLL-14	M	65	67	C	3	9.3	365	248	0	<6	Y	Clb,Alemx4	NR
CLL-15	M	33	35	B	3	14.5	53.5	182	0.67	<12	N	NA	NA
CLL-16	M	47	54	C	3	10.5	289	80	2.43	<12	Y	Clbx2	NR
CLL-17	M	65	74	C	0	10.5	127	58	2.5	<10	Y	Clbx6	PR/24m
CLL-18	M	55	59	C	5	10.6	54.5	212	0	<3	N	NA	NA
CLL-19	M	54	65	B	5	11.7	87	121	8	<12	Y	Clbx6	PR/12m
CLL-20	M	69	70	C	3	9.3	150	26	0.34	<4	Y	Clbx1	NR
CLL-21	F	73	76	C	0	9	171	160	0.42	<10	Y	Clbx6	PR/6 m
CLL-22	F	65	66	B	3	11.8	33	197	4.5	<7	Y	Clbx2	NR
CLL-23	M	77	82	C	-	9.6	362.3	48	0.34	<6	Y	Clbx6	NR
CLL-24	F	62	70	B	3	10.2	81.6	189	4.5	<3	Y	Clbx6	PR/6m
CLL-25	M	65	67	C	4	11.1	298	68	0	<12	N	NA	NA
CLL-26	F	79	86	C	-	10.3	61.4	52	10.4	<12	Y	Clbx6	NR
CLL-27	M	62	76	B	3	11	112.6	204	0	<10	N	NA	NA
CLL-28	M	82	85	B	3	11.6	53.4	185	0	<3	N	NA	NA
CLL-29	M	59	66	C	-	10.2	129.9	63	0.67	<12	Y	Clbx5	NA
CLL-30	M	78	84	C	-	9.4	53.3	165	0	<12	N	NA	NA
CLL-31	F	76	86	C	0	7.7	130	216	0.37	<8	N	NA	NA
CLL-32	F	69	73	C	-	8.9	40.6	142	8.9	<12	N	NA	NA
CLL-33	M	74	79	A	0	14.2	26.8	170	7.07	-	N	NA	NA

^: Indolent and stable CLL. *: F, Female and M, Male. **: Binet stages A, B and C. LDT: Lymphocyte doubling time. * **: Treatment Y, Yes and N, No. Flu: Fludarabine. Clb: Chlorambucil. C: Cyclophosphamide. H: Doxorubicine. O: Vincristine. P: Prednisolone. R: Rituximab. Alem: Alemtuzumab. NR: No response (stable disease). PR: Partial response according to IWCLL (2008). NA: Not applicable. -: Information missed.

3.2.2. Genomic DNA extraction

Genomic DNA from the CLL samples were extracted from the cryopreserved cells and stored in a -20 ^{o}C freezer after measuring the concentration and quality control as described in Section 2.2.2.

3.2.3. DNA concentration measurement and quality control

For measuring the concentration, purity and integrity of starting DNA, respective fluorometric concentration measurement and spectrophotometry and on-chip microfluidic electrophoresis for DNA were used as described in Sections 2.2.3, 2.2.4 and 2.2.5.2.

3.2.4. Target enrichment using HaloPlex technique

Each CLL samples' genomic DNA was digested then was hybridised to the designed HaloPlex probe and PCR amplified using the optimised conditions as explained in Section 2.2.8. Validation of DNA library amplification was performed after using double bead purification of the PCR products of each samples' library as described in Section 2.2.9.

3.2.5. Sequencing template preparation and sequencing on Ion Torrent PGM

Ion Torrent PGM sequencing on Ion 318 chips loaded with equimolar amounts of 4 barcoded DNA libraries were performed as described in Sections 2.2.10, 2.2.11, and 2.2.12. A total of 8 experiments were needed to complete the screening of the cohort comprising the 32 CLL samples.

3.2.6. Sequencing runs assessment, variant calling and sorting

The quality of each sequence run was assessed and recorded by the Torrent Server in a summary statistics report. The quality of raw sequencing data and number of mapped reads were also recorded by the Server. Variant calling was performed for each sample using the optimised custom stringency settings as mentioned in Sections 2.2.13 and 2.2.14. Target region BED file was uploaded to the variant caller software to assign variants only within the

target region, so that variants outside the target regions were discarded. As shown in Figure 3.1, the retained candidate variants were annotated using both Ensembl Variant Effect Predictor and Ion Reporter Software. The variants were compared to available population databases for both somatic mutations (COSMIC-65) and germline mutations (dbSNP-137 and 1000 Genome Project). A group of variants were sorted as SNP and were tabulated for all the samples using excel spread sheet. This was based on their record of validated reference in population based studies namely dbSNP-137 and 1000 genome project. Any identified new variants not reported in the above databases was considered as a SNP, if was found in more than one sample with VAF% of 40 - 60% or 90 - 100% [231]. Similarly, variants with both dbSNP-137 and COSMIC-65 database references were sorted as SNP, if the earlier described conditions existed. Moreover, additional conditions for sorting a variant as somatic mutation were applied including VAF% between 2 - 40% or 60 - 90% and a changeable VAF% identified in its corresponding sample(s) sequenced at earlier stages of the disease for the clonal evolution study in Chapter 4.

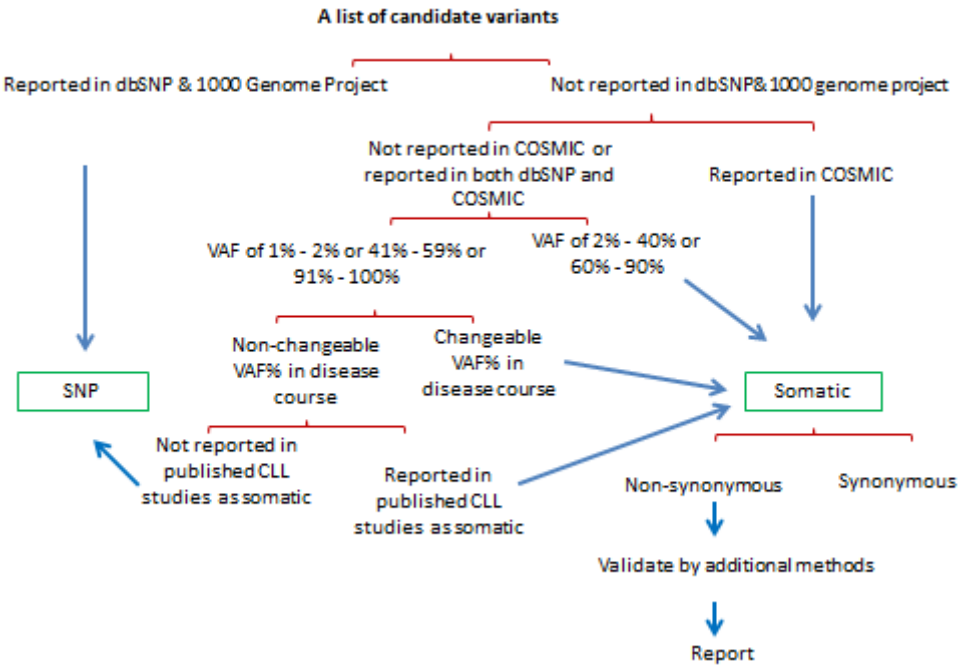


Figure 3.1. **The process of sorting variants called by the optimised TVC and final reporting**

3.2.7. Sanger sequencing for validation of randomly selected variants

Using Sanger sequencing, so far, we have validated 10/22 (45.5%) of the *ATM* and *TP53* mutations detected by the PGM method in the cohort in this chapter as shown in Chapter 2, Sections 2.3.3.2 and 2.3.3.4. It was also important to validate the mutations detected in other genes.

As shown in Table 3.2, various mutations (n = 16) including point mutations (n = 8) and small indels (n = 8) in 11 samples with VAF ranging between 8.5% - 86% in 3 other targeted genes including *NOTCH1, SF3B1* and *PCLO* were selected to be validated by PCR and Sanger sequencing. Those three genes were selected because they were more frequently mutated than other genes (Table 3.6). Therefore, we were able to validate these mutations in multiple samples with the same amplification and Sanger sequencing experiment. The primers required for PCR and sequencing were designed using the NCBI Reference Sequence to amplify regions bearing these mutations as detected with the deep NGS method. The sequence of each primer was compared to the human genome with Blastn Suite (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to exclude any possibility of producing nonspecific sequences.

PCR amplification was set individually for each mutation using the Promega Taq polymerase kit (Promega, UK). All reactions were performed in 0.2-ml PCR tubes in an Eppendorf Mastercycler with standard components, but with individualised DNA amounts and annealing temperature. The details of each mutation and conditions of corresponding PCR are presented in Table 3.2. The specific PCR products were visualised by agarose gel electrophoresis as described in Section 2.2.5.1 to validate successful PCR amplification for each primer set used. Afterwards, 2 - 3 of 50-µl PCR reactions were performed for each sample using the same corresponding PCR conditions to amplify sufficient amounts of PCR products to be used as a template for Sanger sequencing. Following the electrophoresis and visualisation of the bands, the specific bands were excised and purified using the Promega Wizard SV gel and PCR purification kit.

According to the manufacturer's instructions, up to 600 mg of gel slice was melted in 600 µl of membrane binding solution in a 1.5-ml tube with continuous shaking for 10 minutes at 65^{o}C. The tube was spun at 14000 rpm for 1 minute after incubation with a DNA binding column at room temperature for 1 minute. The column-bound DNA was washed in 1200 µl of membrane wash solution and centrifuged for 6 minutes at 14000 rpm. The DNA was then eluted from the column using 40 µl of nuclease free water. The purified products were quantified and sent to Lark Technologies Inc. for bi-directional sequencing using the corresponding primers. Sequences were analysed with Chromas Lite v2.1.1 (Technelysium Pty Ltd, Queensland; Australia) and compared to published wild type NCBI Reference Sequence for the corresponding gene.

Sanger sequencing was provided as a service by Beckman Coulter Genomics (Stortford, UK). The forward and reverse PCR primers were used for the sequencing from both directions.

Table 3.2. **Genes, primer sequences and PCR conditions used to validate various somatic non-synonymous mutations detected by the PGM in different CLL cases**

Gene	Primer name	Primer sequence (5'-3') and the NCBI reference sequence accession number	PCR product size bp	Mutation(s) to be validated Hg19	CLL sample code and (VAF% by PGM)	PCR components 50 µl in 0.2 ml thin walled PCR tubes						Amplification condition*		
						Each Primer pmole	$MgCl_2$ mM	dNTP mM	Gotaq U	g.DNA ng	5x buffer µl	Temperature °C	Time sec.	No. of cycles
NOTCH1	NOTCH1-For NOTCH1-Rev	GGCGGTGCACACTATTCTGC CATCCACAGAGCGCACACAG (XM-011518717.1)	385	9:139390649-50 del CT	CLL-15 (8.5%) CLL-10 (11.3%) CLL-27 (12.4%) CLL-28 (24.0%) CLL-9 (86.0%)	20	75	10	1.25	100	10	94 60 72	30 30 30	32
SF3B1	SF3B1A-For SF3B1A-Rev	TATTACCAACTCATGACTGTCC CATAGTAAGACCCTGTCTCC (NC-018913.2)	374	2:198267491 C>A 2:198267385 A>G 2:198267370 T>C	CLL-5 (47.4%) CLL-4 (49.8%) CLL-20 (80.2%)	20	75	10	1.25	120	10	94 56 72	30 30 30	32
	SF3B1B-For SF3B1B-Rev	TCTGGATGATATTGTGTAACTTAG CTCATCAGGAGACTGGAATTC (NC-018913.2)	430	2:198266834 T>C 2:198266611 C>T	CLL-12 (29.9%) CLL-6 (52.8%) CLL-21 (40.6%)	20	75	10	1.25	100	10			
PCLO	PCLO-For PCLO-Rev	CCAACTCCTGGAAAACCTCC CCTTAGCTGGAGACTGTAGC (NM-033026.5)	376	7:82784832 C>G 7:82784834-35 Ins 30	CLL-5 (36.0 %) CLL-14 (37.2%) CLL-5 (23.5%) CLL-25 (38.2%) CLL-14 (41.4%)	20	75	10	1.25	100	10	94 58 72	30 30 30	32

*The initial denaturation temperature of 94 °C for 3 minutes and the final extension temperature of 72 °C for 5 minutes were used in common for all the PCR reactions

3.2.8. High resolution SNP microarray for identification of chromosomal copy number aberrations (CNA)

In collaboration with Merseyside and Cheshire Regional Genetic Laboratory, we planned to study CNA in samples from 30 CLL cases of the cohort who still had enough (>2µg) DNA available. So far, DNA samples from 14 cases have been tested with CytoSNP 850K array (Illumia, UK). Samples from the remaining 16 cases will be studied in due course. In these samples, 8 recurrent copy number alterations identified by FISH were used as quality control of the SNP array.

The basis of this array technique is Infinium assay as described in Chapter 1 (Section 1.3.1.2). Briefly, DNA samples were amplified, digested and then hybridised to the CytoSNP-850K BeadChip according to manufacturer's instructions. Next, the Chips were scanned using the NextSeq 550 System (Illumina, UK) and raw data files were produced that were further processed using Bluefuse Multi v4.3 (Illumina, UK) software. Log R ratios (measure of signal intensity) were used to identify copy number changes while frequencies of B allele (measure of genotype) were used to identify loss of heterozygosity.

All SNP array analyses (including quality assessment) were performed using the same software. Only CNAs ≥ 5 Mbp and loss of heterozygous state (LOH) ≥ 10 Mbp were included in the downstream analysis [268]. Germline copy number alterations were excluded based on comparison with common polymorphisms listed in the Database of Genomic variants (http://dgv.tcag.ca/) [269]. Cancer genes within regions of both copy number change and loss of heterozygosity were identified based on disease gene status within Online Mendelian Inheritance in Man (OMIM) [270], followed by annotation in the Atlas of Genetics and Cytogenetics in Oncology and Haematology (http://AtlasGeneticsOncology.org/) [271].

3.2.9. Statistical analysis

All statistical analyses were conducted using IBM SPSS v21. Mann- Whitney and Chi square tests were used to assess significance of difference in numbers of mutation events and number of mutated genes between patients with or without chemotherapy prior to sampling. Dominance of mutated genes occurred either as a sole event or co-occurred with

other gene mutations was tasted using Chi square test. Statistical significance was defined as P <0.05. All P values were double sided.

3.3. Results

3.3.1. Good quality of NGS sequencing runs for the CLL cohort

The sequencing data successfully passed all the quality control filters including FASTQC, mapping quality control and the variant calling quality control. As shown in Table 3.3, the coverage analysis revealed homogenously average coverage depth (2252 x) for all of the 33 sequenced samples which was slightly higher than the estimated (1846 x). In detail, almost all (99.98%) of the 52-Kbp target regions were covered by at least one read, with 98.89% by ≥ 20 reads and 96.92% by ≥ 100 reads. Although only 38% of the sequenceable sequences were on targets, the average on-target base reads reached to 60.82%. Furthermore, coverage depth for each target gene of each sample was identified and reported. As shown in Figure 3.2, a high mean coverage depth > 1600 x was achieved for 93.3% (14/15) of the target genes.

Table 3.3. **High level of coverage quality in the NGS sequencing of the study cohort**

Average coverage depth	2252 x
Average uniformity of coverage	92.23%
Average target base coverage at 1 x 20 x 100 x	 99.98% 98.89% 96.92%
Average % of base reads on target	60.82%

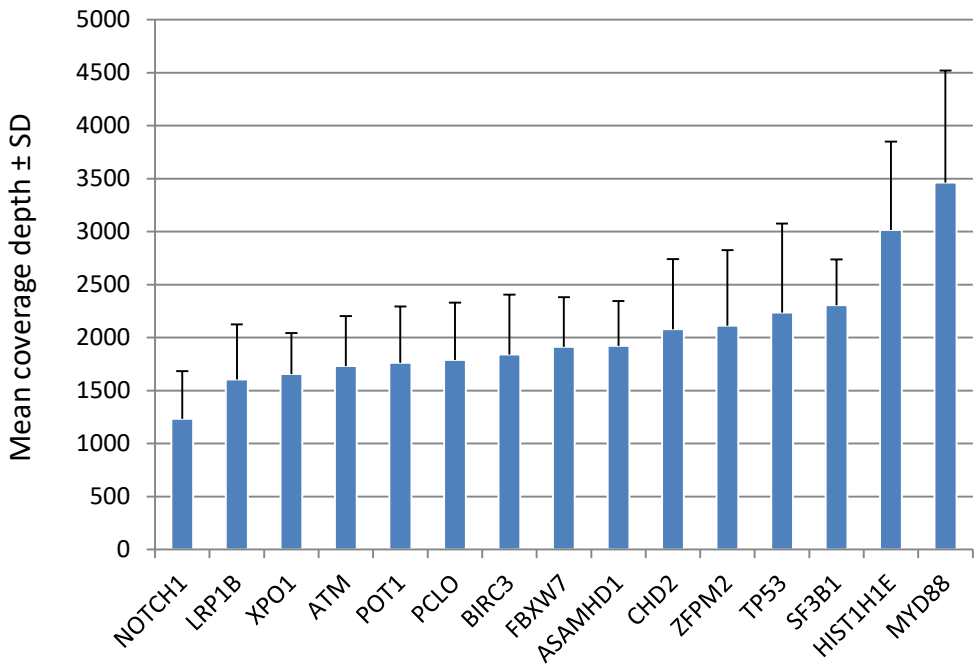


Figure 3.2. **Mean coverage depth per gene**

High coverage depth achieved for the targeted regions of specific genes. *MYD88* had the highest coverage depth while *NOTCH1* had the lowest.

3.3.2. Implementation and fulfilment of criteria set to categorize somatic mutations in the cohort

Using the data analysis setting optimised in Chapter 2, the total number of called variants within the target regions was 812 in the 33 CLL samples analysed (Figure 3.3). As there was no available non-tumour germline material for the cases to be compared to distinguish germline mutations from somatic mutations, we implemented the methods being used in previously published works to separate somatic mutations as mentioned in Section 3.2.6. This enabled us to identify 79/812 somatic non-synonymous mutations that fulfilled at least 2 of the criteria set to be sorted as somatic including being reported in COSMIC-65 database, VAF% of 2-40% or 60-90% [272] and change in VAF% when earlier samples were sequenced. As shown in Table 3.4, the median VAF% change at follow up was 66.66% (range: 0.26% - 2072.72%). All in all, the 79 somatic non-synonymous mutations (mutation events) occurred in 28 patients, with median being 2.0 (range: 1 - 8)/sample. The sample CLL-33 from a case

with stable disease was used as control. As expected, no non-synonymous somatic mutations were detected in it. Notably, 4 other patients with progressive and/or therapy resistant disease had none of these types of mutations detected. Among the 724 germline alterations, only one mutation was not previously reported and did not meet the criteria for somatic mutations as shown in Figure 3.1.

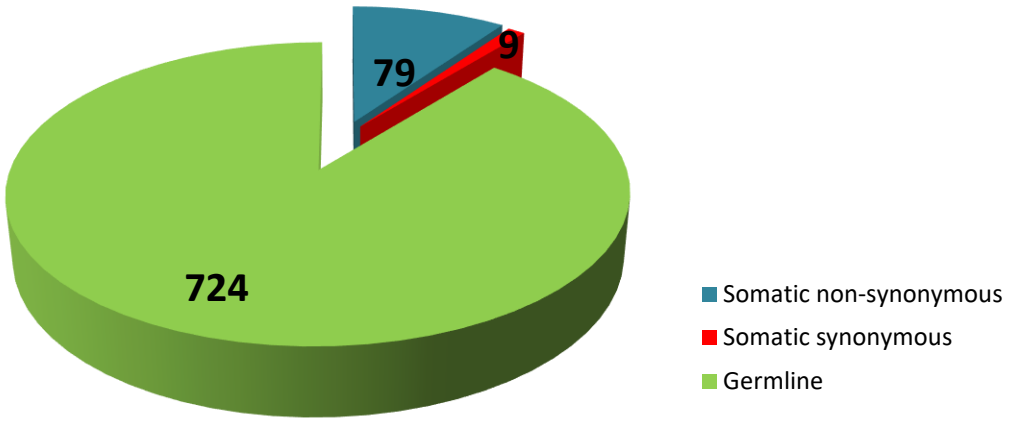


Figure 3.3. **Types and proportions of variants identified within the target gene regions in the 33 CLL samples**

3.3.3. Genomic fingerprints across the cohort ruled out the possibility of sample cross contamination and SNP bias

Similar to Section 2.3.2.1 in Chapter 2, the SNP patterns of this cohort were used to exclude cross sample contamination. These patterns clearly showed a unique fingerprint of each sample (Table 3.5). This provided good evidence not only for the unbiased SNP calling, but also non-contamination between samples in the tests. Moreover, we calculated the total number of non-synonymous germline mutations for the whole cohort and the indolent CLL

case, a mean ± SD of 8.45 ± 2.1 of non-synonymous SNPs reported in dbSNP-137 data base was identified/sample as shown in Table 3.5.

3.3.4. Validation of candidate somatic non-synonymous variants by additional methods

3.3.4.1. Validation of 9:139390649-50 deletion CT in *NOTCH1*

NOTCH1 is a gene frequently mutated in patients with CLL and Richter's transformation. Exon 34 is the only exon found to be affected by mutations in CLL. The most common mutation reported in this disease is a deletion of a dinucleotide which is located at the 3' end of the coding sequence of this exon. Shallow coverage over a stretch of 300 bp at this end of the coding sequence is frequently encountered. This is because this region is rich in G or C nucleotides. In our cohort, the most common mutation repeatedly found in multiple samples was this hotspot. To exclude possibility of sequencing error particularly for an indel in *NOTCH1* which had shallower coverage depth (Mean: 1235 x) (Figure 3.2) relative to the high coverage depth for all genes (Av: 2252 x) (Table 3.3), bidirectional Sanger sequencing method was used to validate the mutation status of 7 CLL cases from the cohort who had been identified as wild type (Wt., n = 2) and as mutated (Mu., n = 5) by the deep NGS. The mutated samples had different variant allele frequencies ranging from as low as 8.5% to 86%. The wild type samples were chosen to be used as controls. The results showed a high agreement of the location and VAF between the two sequencing methods (Figure 3.4).

Table 3.4. **Criteria set for somatic non-synonymous mutations**

Variant information					Being reported		Not being reported	VAF% categories	VAF Changes at followup	
Gene	Chr.location (hg19)	Ref	Var	Amino-acid change	COSMIC 65	dbSNP 137		2-40% or 60-90%	% of change	Comments
TP53	17:7572969	8N	-	p.HPPX380X	Yes			Yes	1425%	
TP53	17:7576854	T	C	p.Q331R	Yes			No	11.14%	
TP53	17:7576891	T	A	p.K319X	Yes			Yes	257.14%	
TP53	17:7577079	C	A	p.E278X	Yes			Yes	233.30%	
TP53	17:7577100	T	C	p.R280G	Yes			Yes	0.26%	Short follow-up interval
TP53	17:7577118	C	G	p.V274L	Yes			Yes	100%	
TP53	17:7577120	C	T	p.R273H	Yes	Yes		Yes		No follow-up
TP53	17:7577538	C	T	p.R248Q	Yes	Yes		Yes	29.01%	
TP53	17:7577551	C	T	p.G244S	Yes			yes	161.07%	
TP53	17:7577567	A	C	p.C238W	Yes	Yes		Yes		No follow-up
TP53	17:7577568	C	T	p.C238Y	Yes			No	416%	
TP53	17:7577580	T	C	p.Y234C	Yes			Yes	200%	
TP53	17:7578211	C	A	p.R213L	Yes			No	296.35%	
TP53	17:7578394	T	C	p.H179R	Yes			No	2072%	
TP53	17:7578413	C	A	P.V173L	Yes			Yes	451%	
TP53	17:7578525	G	C	p.C135W	Yes			Yes	98.71%	
ATM	11:108106443	T	A	p.D126E	Yes	Yes		No	5.50%	Reached clonal level
ATM	11:108121763	G	A	p.W524X			Yes	No	8.60%	
ATM	11:108143528	T	G	p.L1078R			Yes	Yes	23.12%	
ATM	11:108168106-107	-	C	p.L1668fs			Yes	Yes	13.38%	
ATM	11:108175530-31	-	T	p.H1876fs			Yes	yes	66.66%	
ATM	11:108183216	A	-	p.S2000fs			Yes	yes		No follow-up
ATM	11:108186598	T	C	p.Y2019H	Yes			No	1.90%	Reached clonal level
ATM	11:108186599	A	G	p.Y2019C	Yes			No	23.9	
ATM	11:108186757	G	A	p.E2039K	Yes			No		No follow-up
ATM	11:108198445-54	10N	-	p.E2351fs	Yes			No	62.90%	
ATM	11:108213973	G	A	p.G2765S	Yes			Yes	1.86%	
ATM	11:108216487	T	-	p.S2812X			Yes	Yes	500%	
ATM	11:108216582-27	5N	-	p.W2845fs			Yes	No	5%	
ATM	11:108236179	G	C	p.A3039P			Yes	Yes	660%	

Table 3.4. **Criteria set for somatic non-synonymous mutations** (continued)

Variant information					Being reported		Not being reported	VAF% categories	VAF Changes at followup	
Gene	Chr.location (hg19)	Ref	Var	Amino-acid change	COSMIC 65	dbSNP 137		2-40% or 60-90%	% of change	Comments
SF3B1	2:198266611	C	T	p.G742D	Yes			No	83.71%	
SF3B1	2:198266834	T	C	p.K700E	Yes			Yes	199%	
SF3B1	2:198267360	T	A	p.K666M	Yes	Yes		Yes	250%	
SF3B1	2:198267370	T	C	p.T663A			Yes	Yes	46.08%	
SF3B1	2:198267385	A	G	p.W658R			Yes	No	0.80%	Short follow-up interval
SF3B1	2:198267491	C	A	p.E622D	Yes			No	17.32%	
SF3B1	2:198267699	G	A	p.R594X			Yes	Yes	490%	
SF3B1	2:198268383	G	A	p.R549C	Yes			Yes	320%	
SF3B1	2:198269834	A	T	p.L502X			Yes	Yes	300%	
PCLO	7:82390770	T	C	p.K5016R			Yes	No	54.56%	
PCLO	7:82508679	G	A	p.A4543V			Yes	Yes	305%	
PCLO	7:82581607	C	T	p.V2888I	Yes			No	161.07%	
PCLO	7:82581658	G	T	p.P2871T				Yes	590%	
PCLO	7:82583008	G	A	p.P2421S			Yes	Yes		No follow-up
PCLO	7:82784832	C	G	p.Q375H	Yes	Yes		Yes	3.76%	
PCLO	7:82784834-35	-	30N	p.Q374_Q375Ins10			Yes	Yes	8.12%	
LRP1B	2:141032021	T	A	p.N4372Y	Yes	Yes		No	17.50%	
LRP1B	2:141032152	T	A	p.Y4328F			Yes	Yes	380%	
LRP1B	2:141259338	C	T	p.R2923K			Yes	Yes	590%	
SAMHD1	20:35526319	C	T	p.R551Q			Yes	No	0.60%	Reached clonal level
SAMHD1	20:35526885-86	-	A	p.K523fs*			Yes	Yes	52.50%	
SAMHD1	20:35545207	T	-	p.N327fs			Yes	Yes	210%	
SAMHD1	20:35580045	A	T	p.M1K			Yes	Yes	23.96%	
XPO1	2:61719472	C	T	p.E571K	Yes			Yes	50.00%	
FBXW7	4:153247330	C	G	p.G491A			Yes	Yes	980%	
FBXW7	4:153249384	C	A	p.R465L	Yes			Yes	20.45%	
HIST1H1E	6:26156965	C	T	p.A116 V			Yes	No	22.37%	
NOTCH1	9:139390649-50	CT	-	p.F2482Ffs*2			Yes	Yes	50%	CLL hot spot
NOTCH1	9:139390945	G	A	p.Q2416X	Yes			Yes	491.42	
BIRC3	11:102207709	A	-	p.V565fs			Yes	Yes	827%	
CHD2	15:93499738	A	T	p.H620L	Yes			Yes	29.57%	
CHD2	15:93499735	C	G	p.A619G			Yes	Yes	38.49%	

Table 3.5. **SNPs being reported in dbSNP-137 and identified within the target region in the cohort of 33 CLL cases***

Chr.location (hg19)	Ref	Var	Class	dbSNP	CLL-1	CLL-2	CLL-3	CLL-4	CLL-5	CLL-6	CLL-7	CLL-8	CLL-9	CLL-10	CLL-11	CLL-12	CLL-13	CLL-14	CLL-15	CLL-16	CLL-17	CLL-18	CLL-19	CLL-20	CLL-21	CLL-22	CLL-23	CLL-24	CLL-25	CLL-26	CLL-27	CLL-28	CLL-29	CLL-30	CLL-31	CLL-32	CLL-33
11:102201850	G	A	Missense	rs17881197													52																				51
11:102201948-9	AG	-	Frameshift del	rs151072309																										59							
11:102206908	CT	-	Frameshift del	rs370069893																										59							
11:102207851	G	A	Synonymous	rs1055088	100	50	100				48	50	100	39	46	99	60		100	43		98	48	100	100	100	36	100	100		99	99	100	48	100	100	41
11:108119770	C	G	Synonymous	rs1800727										42																							
11:108121446	A	G	Synonymous	rs4987943										49																							
11:108138003	T	C	Missense	rs1800056			51																		49										56		
11:108143456	C	G	Missense	rs1800057			52																		47										49		
11:108160350	C	T	Missense	rs1800058						54																											
11:108163487	C	T	Synonymous	rs1800889	44								54											50													
11:108170506	A	C	Missense	rs1800059													57																			56
11:108175462	G	A	Missense	rs1801516			48			50												95	55			52						44				
11:108183167	A	G	Missense	rs659243	100	100	100	100	100	100	100	100	99	100	100	100	100	100	100	100	100	99	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
11:108196260	A	C	Missense	rs45481995						31																											
11:108216487	T	-	Frameshift del	rs1800889																			51														
15:93510603	A	G	Synonymous	rs4777755	99	99	98	99	50	51	97	50	98	56	98	98	98	52	99	99	53	99	99	99	55	98	49	52	98	100	51	99	99	50	99	49	99
15:93521604	A	G	Synonymous	rs11074121	100	100	99	100	50	46	98	48	98	52	100	99	99	53	99	99	53	100	99	99	49	99	50	52	98	100	47	99	99	46	100	47	100
15:93536197	C	T	Synonymous	rs2272457	47	48	48	100				50	51	50		47	91		55	48			49	100		53	50		48	49	47	48	46		49	49	
15:93552488	C	T	Synonymous	rs34315566					52											49								51								
17:7578210	T	C	Synonymous	rs1800372																			51								50					
17:7579472	G	C	Missense	rs1042522	97	39	68	97		98	28	99	96		65	49	90	93	62	99		47	48	57	61	99	96		95	98	91	99	98	95	97	98	55
2:141032021	T	A	Missense	rs149644677																					49											
2:141032088	C	T	Synonymous	rs1386356	49	48	47	100	100	100	47		100		50	50	53	49	100		48	49	100	91	100	99	50	50	100	48	48	46	100	49	49	100	51
2:141072519	C	G	Missense	rs17386226		42								55																50							
2:141092084	T	G	Missense	rs79879036		52																															
2:141116420	C	T	Missense	rs150879175																					49												
2:141116447	G	T	Missense	rs35546150											47		54																				53
2:141128779	C	G	Missense	rs76554185															52				46														
2:141130695	C	T	Synonymous	rs16843864		54		45	100				48	48					44				47	90	100					48		49					
2:141232800	C	T	Missense	rs72899872																								50						52	48	

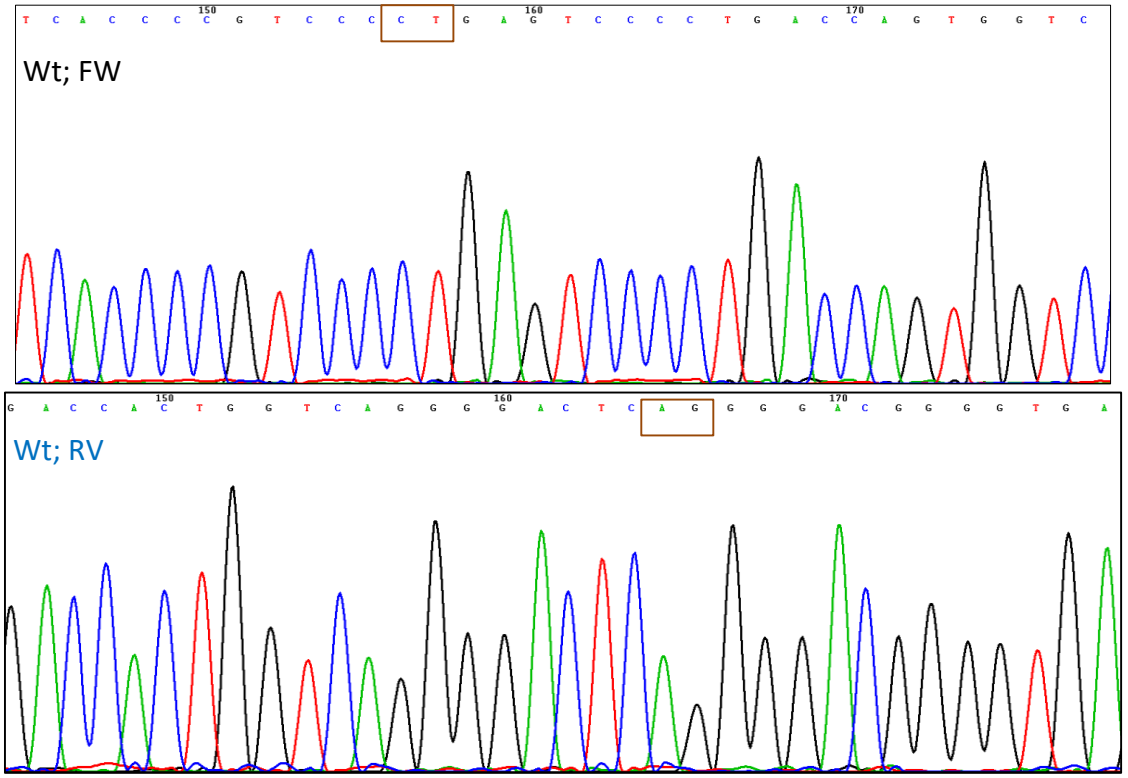
Table 3.5. **SNPs being reported and identified within the target regions in the cohort** (continued)

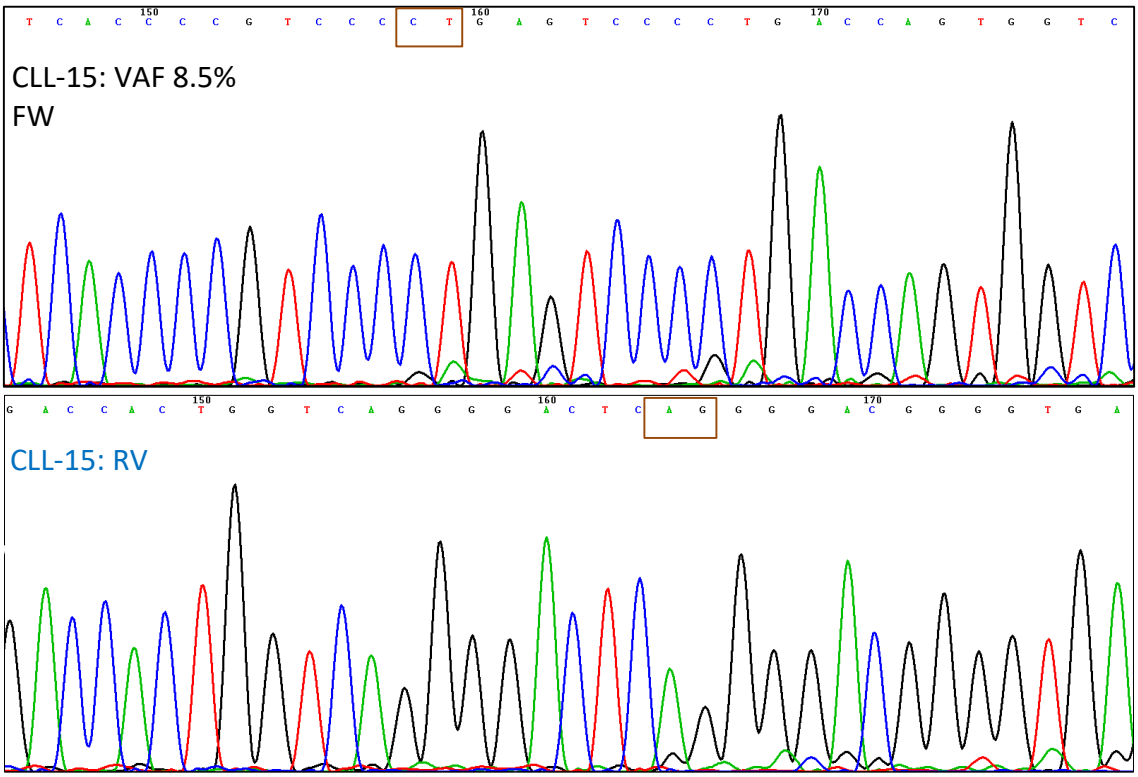
Chr. location (hg19)	Ref	Var	Class	dbSNP	CLL-1	CLL-2	CLL-3	CLL-4	CLL-5	CLL-6	CLL-7	CLL-8	CLL-9	CLL-10	CLL-11	CLL-12	CLL-13	CLL-14	CLL-15	CLL-16	CLL-17	CLL-18	CLL-19	CLL-20	CLL-21	CLL-22	CLL-23	CLL-24	CLL-25	CLL-26	CLL-27	CLL-28	CLL-29	CLL-30	CLL-31	CLL-32	CLL-33
2:141242918	T	C	Missense	rs34488772															52				53														
2:141245204	T	C	Synomymous	rs74789055															46				41														
2:141259283	G	A	Synonymous	rs35296183			48	47					48						47		45	51	46					99	47	47			99		45	99	
2:141260668	A	G	Synonymous	rs4444457	50		50	47		53	52		46		50	50	51	51	100		100	49	100			100	50	100	100	45		47	100	100	100	100	53
2:141272253	C	T	Synonymous	rs61732738										50																						
2:141274576	T	C	Synonymous	rs4954672	50	100	99	98	99	47	51	99	99	53	51	99	94	49	100	98	47	99	100	99	98		98	100	50	99	98	98	53	51	49	97	52
2:141298635	C	T	Missense	rs146867394																				52								53					
2:141457962	C	T	Missense	rs34694228						50																											
2:141457985	T	A	Synonymous	rs13431727			49	48					48	48						43		44		90			45	98	46						51	99	
2:198265526	A	G	Synonymous	rs788018	100	100	100	100	53		100	100	100	53	46	49	55	45		100	100	100	100	100	52	53	52	50		100	100		47	50	46		100
2:198274685	C	T	Missense	rs377192403						45																											
2:61719275	T	C	Synonymous	rs3816341												52													51								
2:61722724	G	T	Synonymous	rs6171632										41																							
3:38182062	T	C	Missense	rs148149492																					39											
6:26157073	A	G	Missense	rs2298090																47					49											
7:124481185	C	A	Missense	rs 35536751							55				48													50								
7:82435033	C	T	Synonymous	rs12668093	54			42			42							50						49		50		52					49	100		51	
7:82451836	G	A	Synonymous	rs146099474				40																												
7:82453708	A	C	Missense	rs2522833	100	47	49	100		53		48	48		51		47	97	50	95				47	96		50	49	47				98	97	49	48	95
7:82544510	C	T	Synonymous	rs61995908																																	
7:82544987	A	G	Synonymous	rs17156844				59	42	53	42				51	51	49			53				52		47	51					50	51		50		
7:82545070	G	A	Missense	rs150669313																				51													
7:82579183	T	C	Missense	rs201013392								44																									
7:82580293	A	T	Missense	rs199626449																									47								
7:82581489-90	-	TCA	Nonframeshift ins	rs10630259										99			98			99					99	99	48			98		99			98	98	
7:82581859	C	T	Missense	rs976714	57		51	57	44	51		46			51	51	56	55	54	99	46			50	100		48		50			52	49		52		100
7:82582258	T	G	Missense	rs10261848										52																						
7:82582846	C	T	Missense	rs10954696	49		51	56	40	52		47			48	46	52	48	50	100	48			50	100		52		51			46	50		47		100
7:82583280	C	T	Missense	rs17148149										49																						
7:82583388	A	G	Missense	rs10487647										51																						

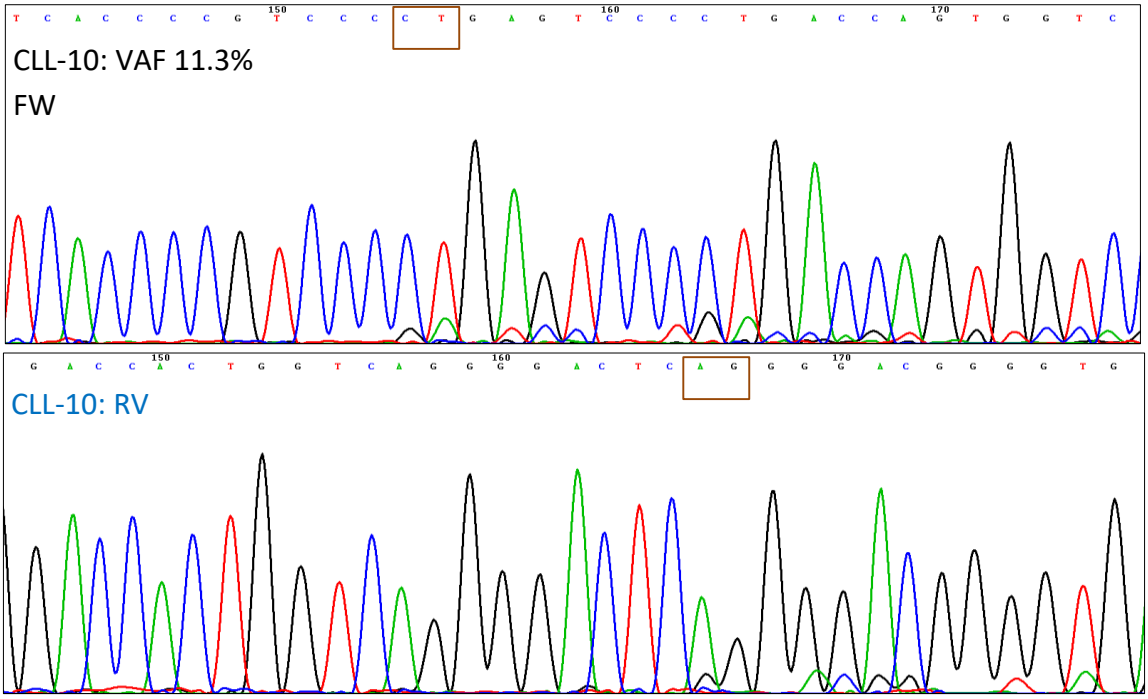
Table 3.5. **SNPs being reported and identified within the target regions in the cohort** (continued)

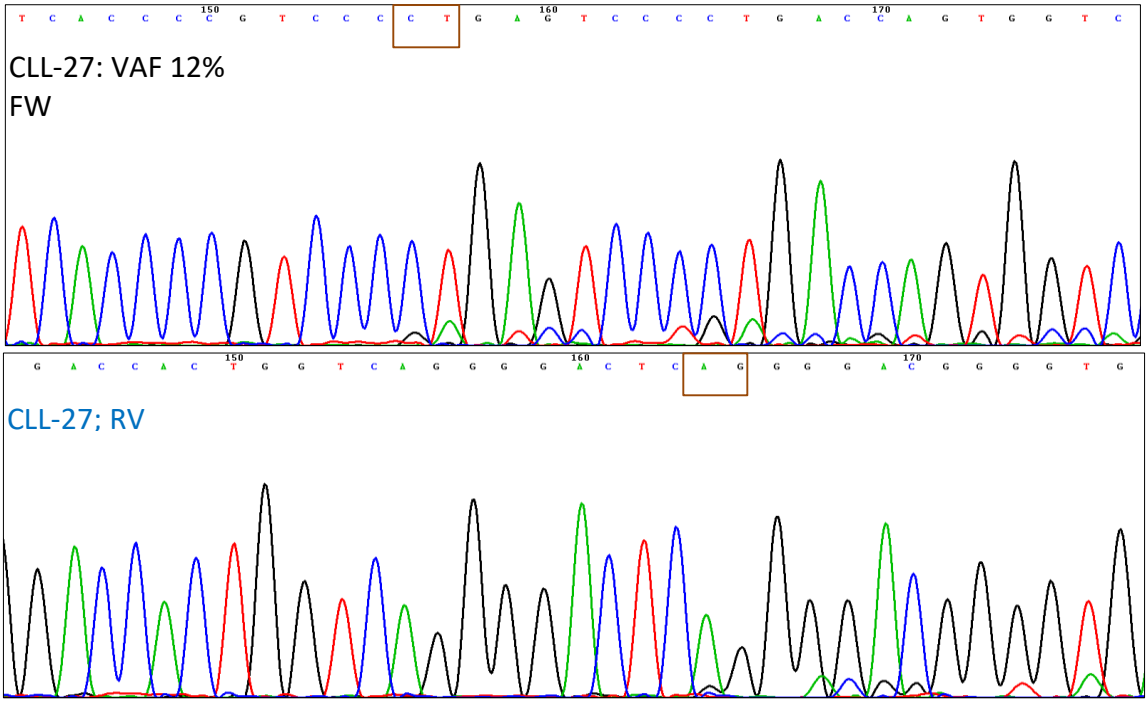
Chr. Location (hg19)	Ref	Var	Class	dbSNP	CLL-1	CLL-2	CLL-3	CLL-4	CLL-5	CLL-6	CLL-7	CLL-8	CLL-9	CLL-10	CLL-11	CLL-12	CLL-13	CLL-14	CLL-15	CLL-16	CLL-17	CLL-18	CLL-19	CLL-20	CLL-21	CLL-22	CLL-23	CLL-24	CLL-25	CLL-26	CLL-27	CLL-28	CLL-29	CLL-30	CLL-31	CLL-32	CLL-33
7:82583609	A	C	Missense	rs10487648	37							47				47		41	44	44					97				40			39					43
7:82584574	G	T	Missense	rs61995911											42																					
7:82585803	G	C	Missense	rs114445550										38																						
7:82595324	C	T	Synonymous	rs9969358										54																						
7:82595742	T	C	Missense	rs28680905										47																						
7:82764425	C	G	Missense	rs2877	58		52	100	100	100	60	47	53				50	100	100	99	55	41		100	100	57	59	39	44	43	47	100	46	99			100
7:82784456	A	G	Missense	rs6972461	47	44	41	38			34	43	41		100	100	44				44	46	100			47	45		40	47	42	46	45		100	44	
7:82784501	G	T	Missense	rs201808333																												53			
7:82785097	T	C	Missense	rs61741659		44							46	50			43		46		45			45	41	40		49		44				46		44	43
7:82785099	T	G	Synonymous	rs61741653									53		48	50													50						54		
7:82785271	G	A	Missense	rs1121444																47																	
7:82785304	G	A	Missense	rs61738783			50		55																												
8:106813518	C	G	Missense	rs11993776	49									100	100				50								49									50	
8:106813672	A	G	Synonymous	rs920628										51	66																					
8:106814086	T	C	Synonymous	rs16873732											64																					
8:106814279	A	G	Missense	rs28374544										52																						
8:106814656	G	C	Missense	rs2920048		51										52				99	53								51		47			51			
8:106814695	C	G	Synonymous	rs355998713											67																					
8:106815286	T	C	Synonymous	rs1442320											62																					
8:106815474	C	T	Missense	rs16873741										41																						
8:106815679	A	G	Synonymous	rs16873744										100																						
9:139390958	T	C	Synonymous	rs11574911										42	45																					
9:139391636	G	A	Synonymous	rs2229974			51	50	100	100	53	100	100	58	50	53	52	51	100	52		100	53	100	53	51	100			100	100		53	52		100	53
Total No. of non-synonymous SNPs					9	8	11	7	5	11	5	9	6	12	10	7	12	7	11	11	7	5	8	9	8	9	9	6	10	9	5	10	7	7	10	8	11

* The distinct SNP information including chromosomal location using Hg19 as reference, nucleotide change, effect on the amino-acid and dbSNP-137 reference numbers are shown (rows). The SNP status for each sample (columns) are marked as heterozygous (VAF 40% - 60%) by green, homozygous (VAF 90% - 100%) by red and no call (VAF 0%) by blank.


Wt; FW
Wt; RV


CLL-15: VAF 8.5%
FW
CLL-15: RV


CLL-10: VAF 11.3%
FW
CLL-10: RV


CLL-27: VAF 12%
FW
CLL-27; RV

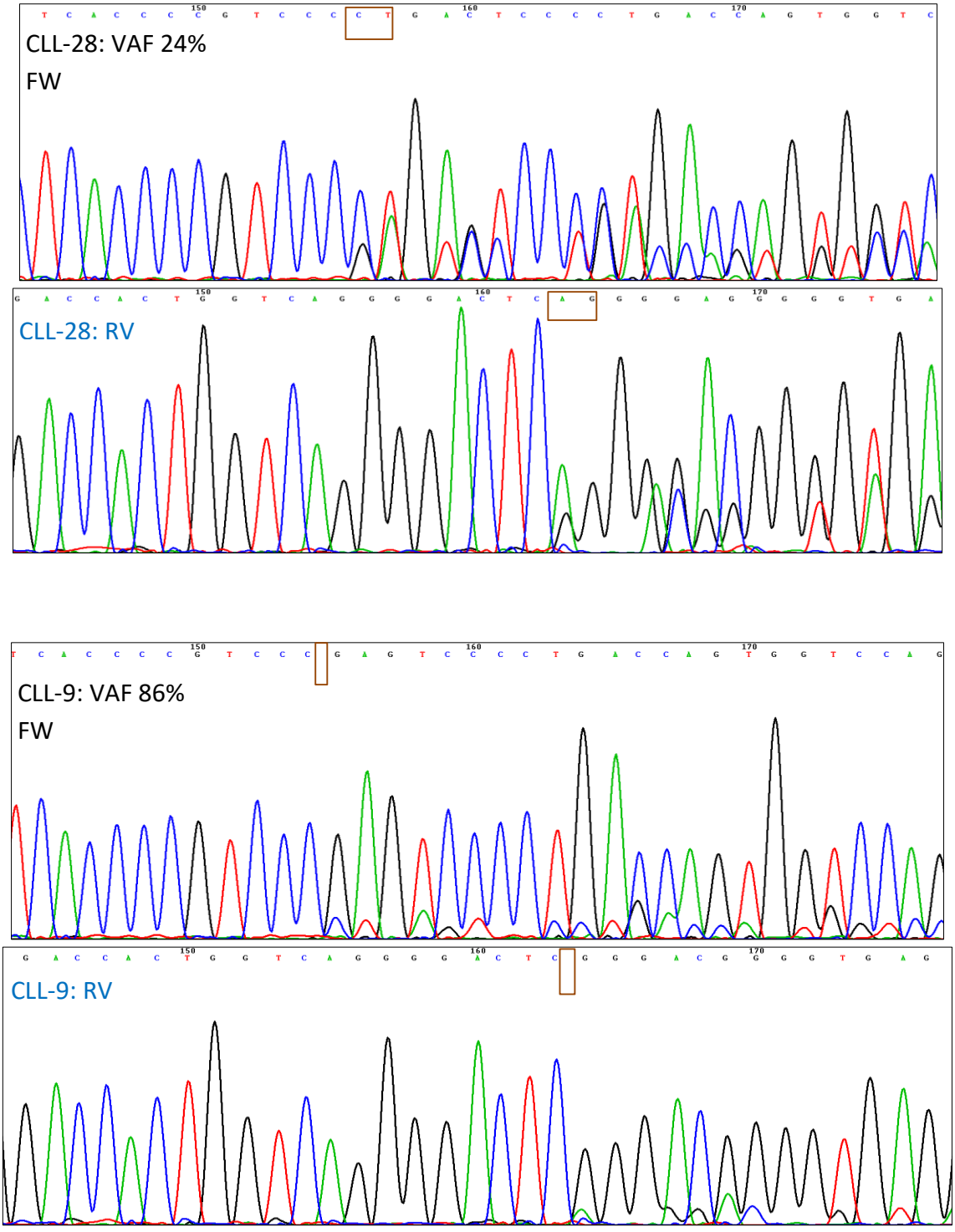
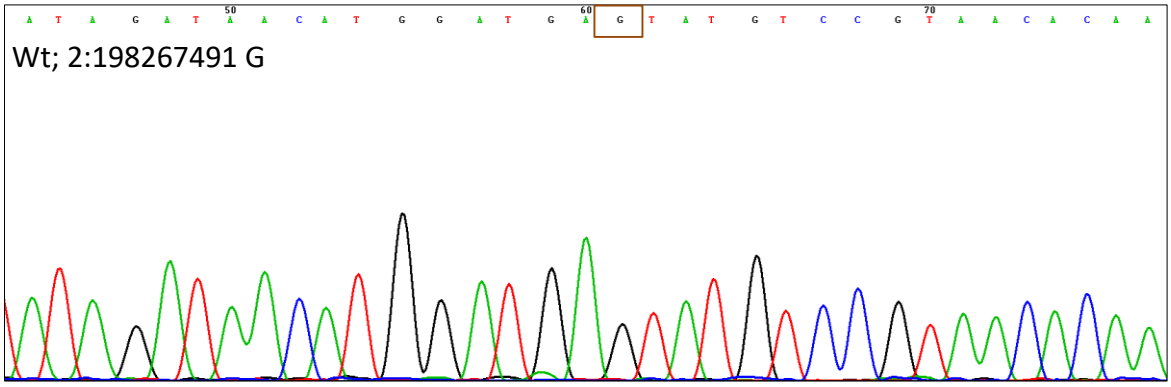


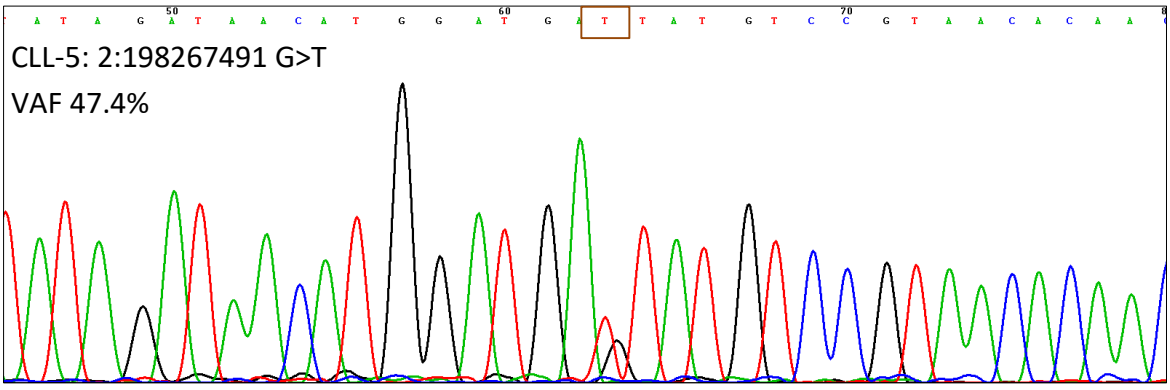
Figure 3.4. ***NOTCH1*, 9:139390649-50 del CT validation using bidirectional Sanger sequencing**

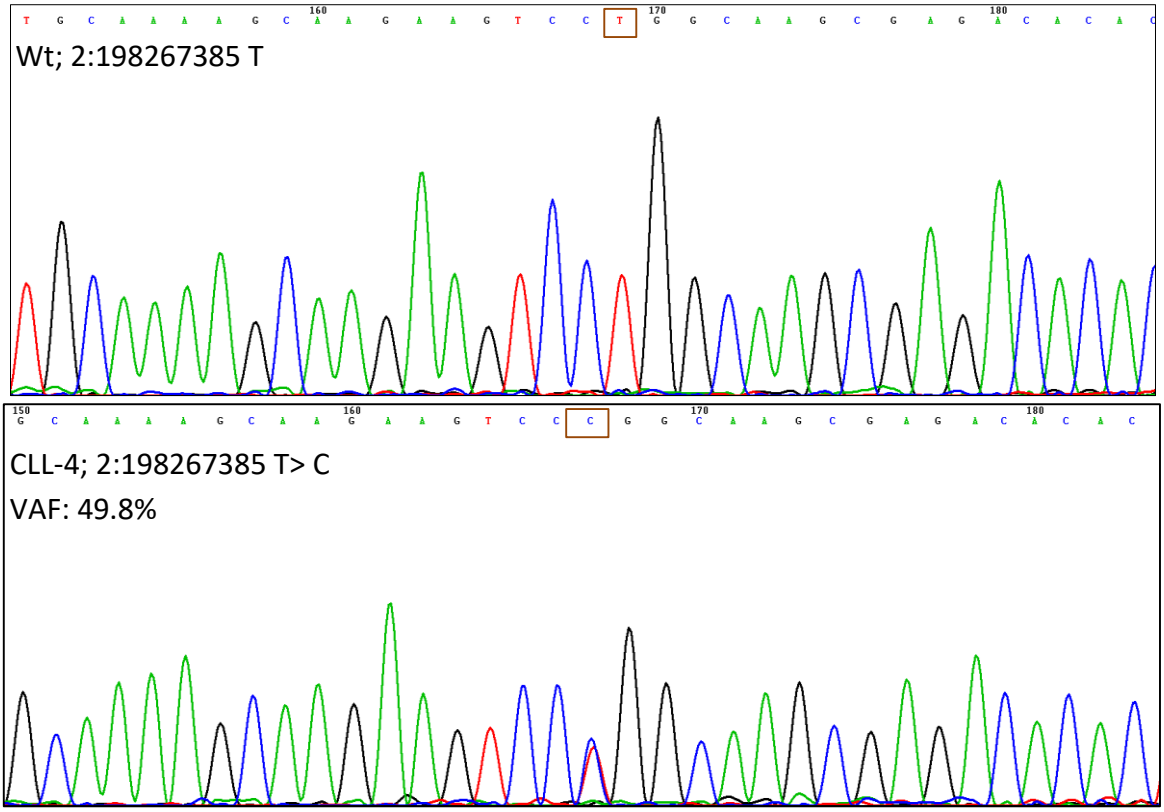
The forward (FW) and the reverse (RV) Sanger sequencing chromatograph is illustrated for a wild type (Wt) and the 5 mutated (Mu) cases marked with their IDs. The different levels of variant allele frequencies VAF % identified by the Ion Torrent PGM are shown to demonstrate the high concordance between the VAF % and the chromatograph peaks of the mismatched bases. The dinucleotide (CT) deleted is enclosed with the small box.

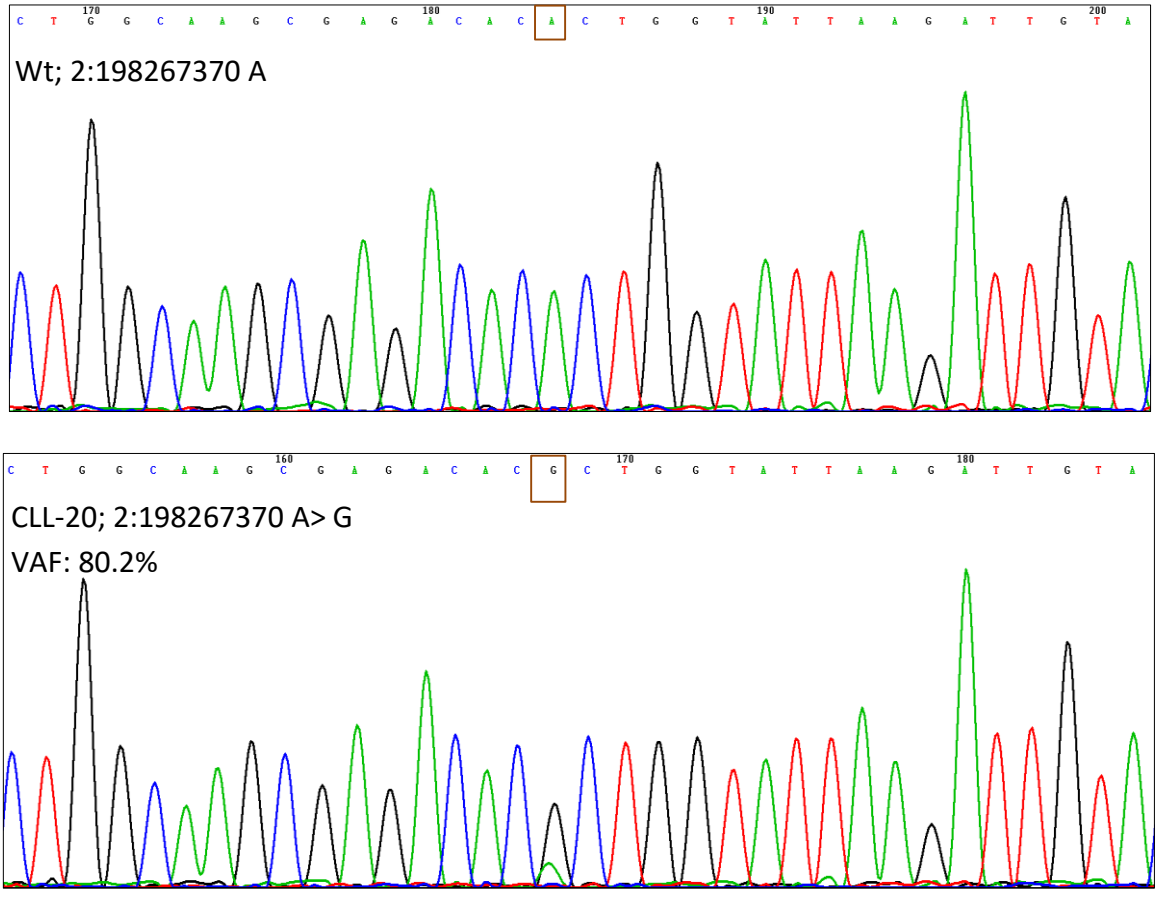
3.3.4.2. Validation of mutations in *SF3B1*

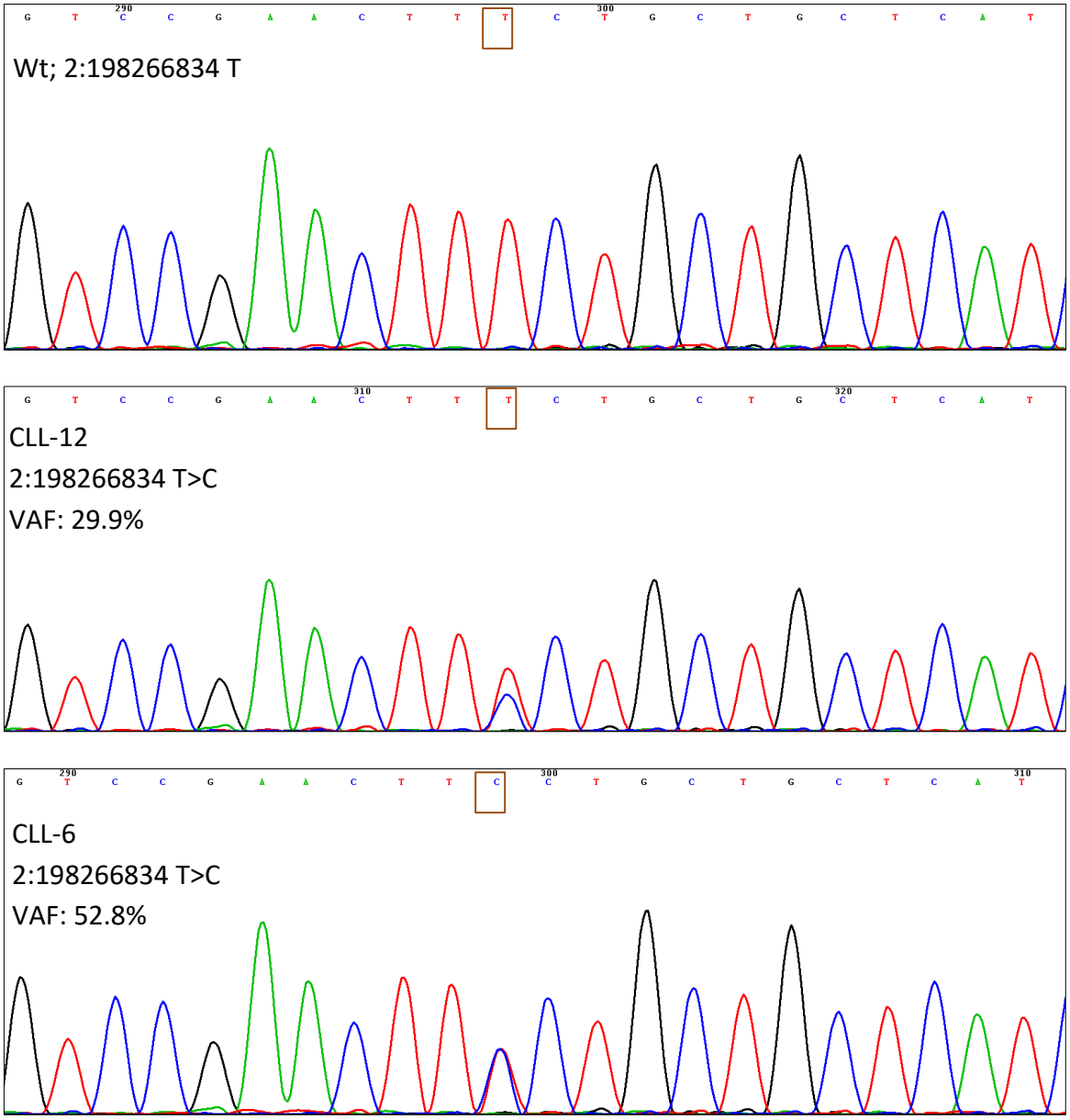
This gene was the most frequently mutated gene in this study. We selected various mutations in this gene to be validated by bidirectional Sanger sequencing as described in Section 3.2.7. The mutated samples as well as wild type control samples for these chromosomal locations were subjected to PCR amplification using the primer pair and PCR conditions described in Table 3.2. There was high concordance between the Sanger sequencing and the NGS results as shown in Figure 3.5.






Wt; 2:198267385 T
CLL-4; 2:198267385 T> C
VAF: 49.8%


Wt; 2:198267370 A
CLL-20; 2:198267370 A> G
VAF: 80.2%


Wt; 2:198266834 T
CLL-12
2:198266834 T>C
VAF: 29.9%
CLL-6
2:198266834 T>C
VAF: 52.8%

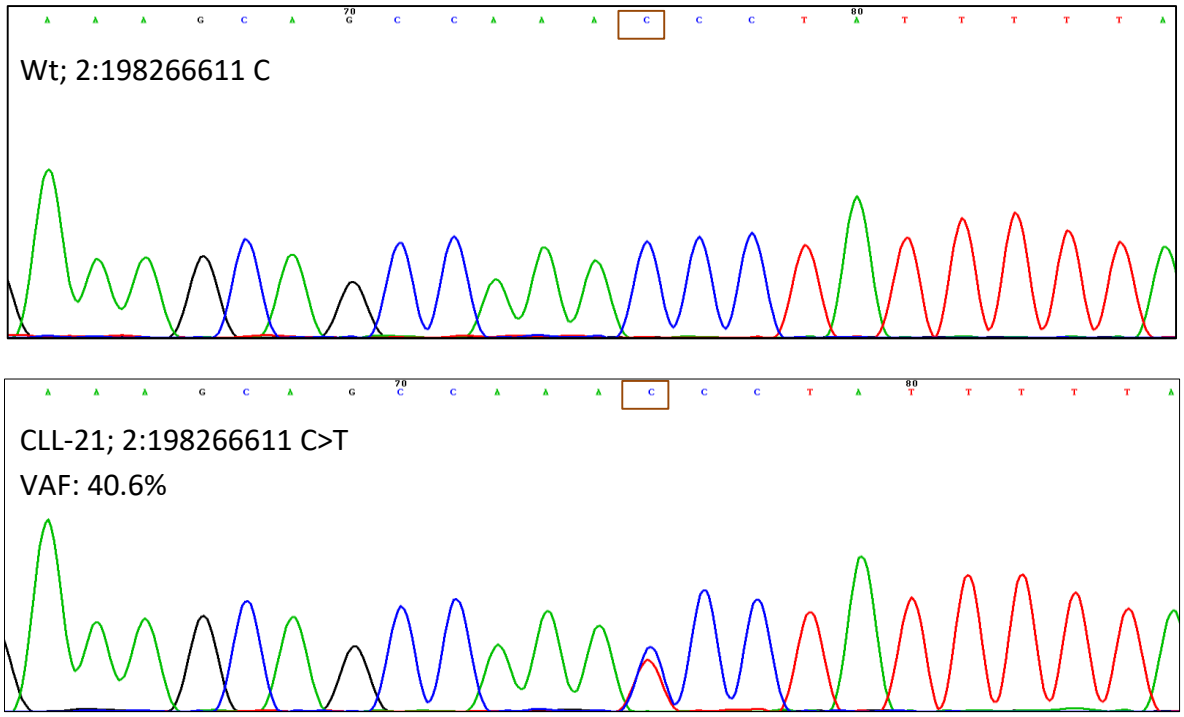


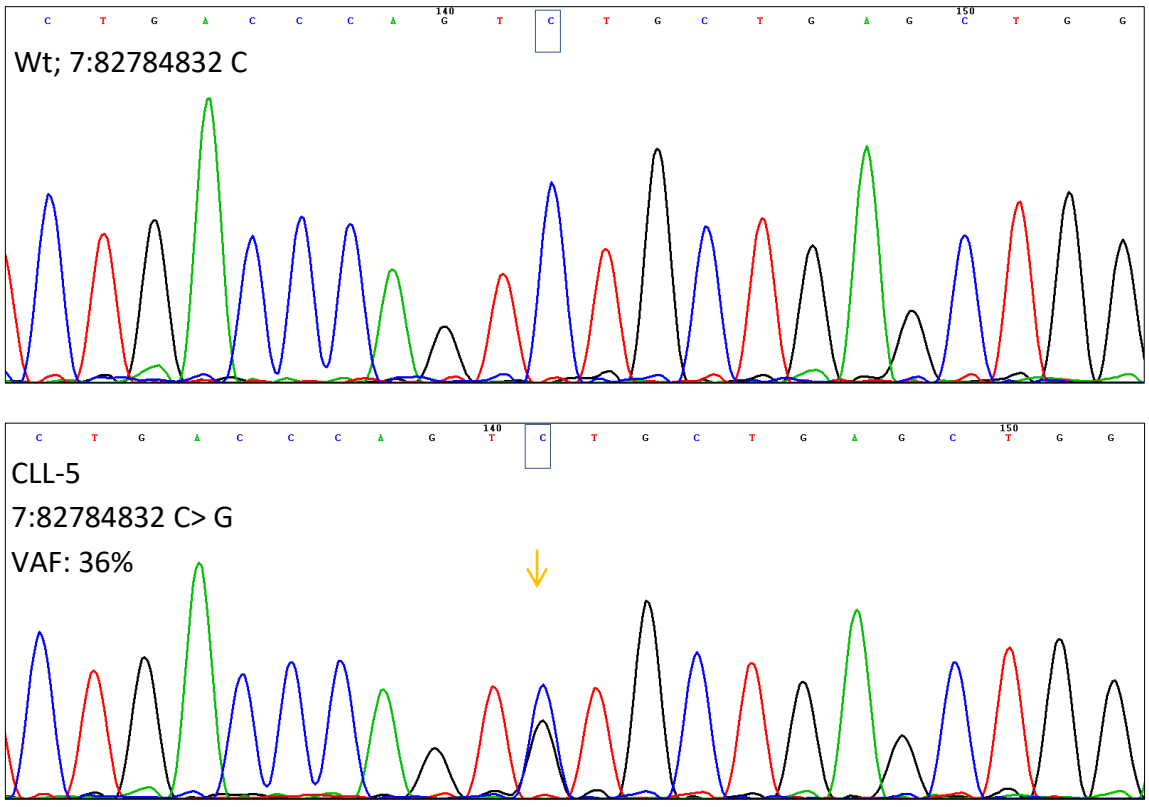
Figure 3.5. **Sanger sequencing for validation of various mutations in *SF3B1***

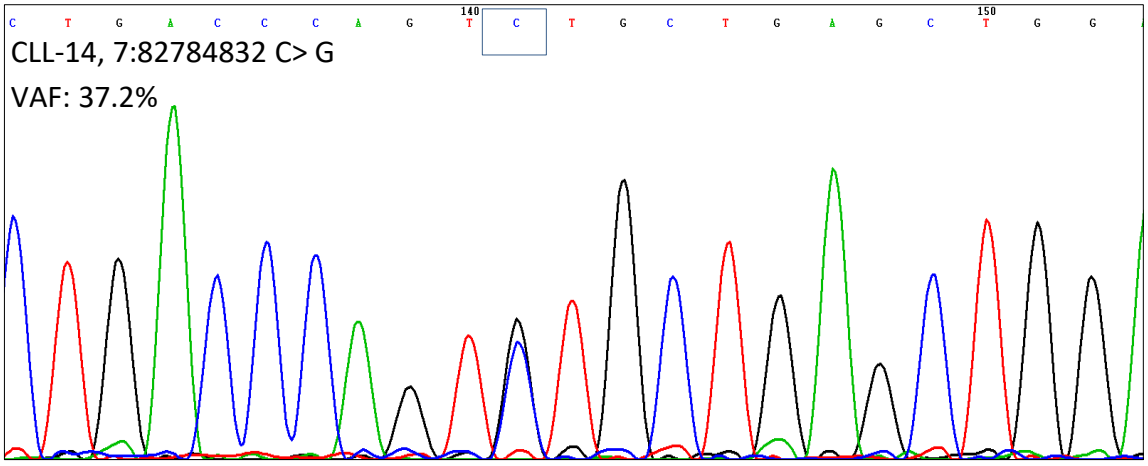
The chromatograph of 6 CLL cases marked by their identification code, the *SF3B1* point mutation and the variant allele frequency (VAF %) along the corresponding wild type and reference allele for each mutation are demonstrated (only chromatograph of sequencing from forward direction is shown for each mutation in each sample). The results were highly consistent with the Ion Torrent PGM mutational status assigned for each case. The nucleotides enclosed in small boxes denote the reference in the wild type sample (Wt) and the variant allele in the corresponding mutated sample(s).

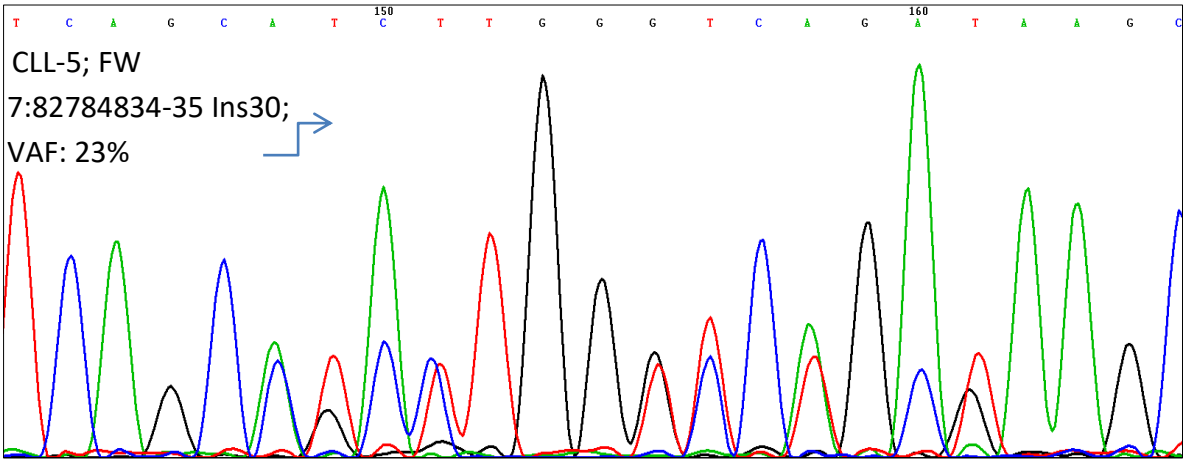
3.3.4.3. Validation of mutations in *PCLO* by Sanger sequencing

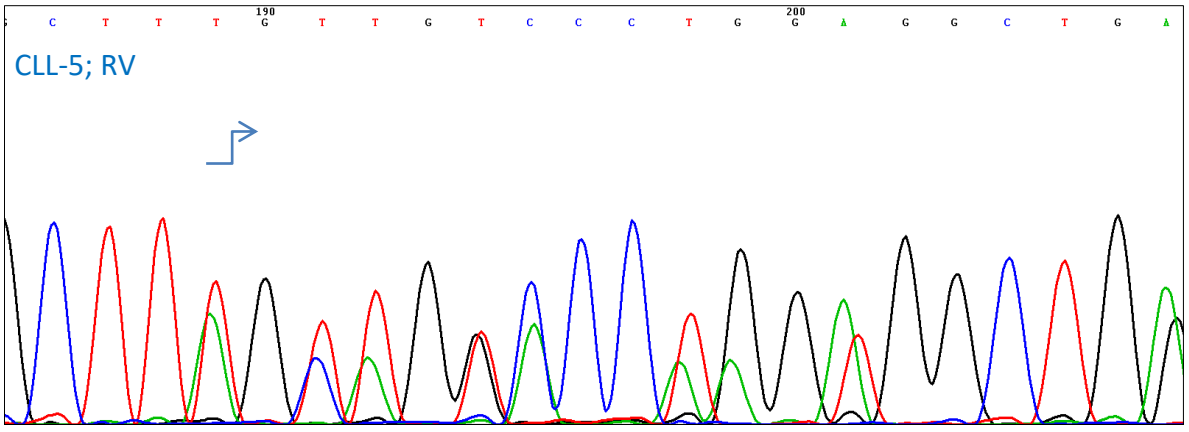
This gene was among the most frequently mutated gene in this study. We selected various mutations including a point mutation and a 30 bp insertion to be confirmed. A sequence gap between the point mutation and the insertion of only 1 nucleotide was detected by the PGM. 2 CLL samples carried both changes but with different allele frequencies while another CLL case had the same insertion without the point mutation (as shown in Table 3.2). PCR amplification and subsequent Sanger sequencing was performed in both forward (FW) and reverse (RV) directions to validate the exact location and the size of the insert. According to

the Ion Torrent PGM result, the 30bp sequence of insert was **CTCTTGGTCCTGCTAAGCCTCCAGCTCAGC**. The bidirectional Sanger sequencing confirmed both the single nucleotide changes in the 2 samples identified by NGS. While Sanger sequencing for the 30 bp insertion in each of the three samples were consistent to the NGS result only in the forward sequencing direction and not in the reverse direction. The variation figured as a gap of 30 bp between the point mutation and the insertion in samples harboured both variants rather than the 1 bp gap identified by the PGM (Figure 3.6). Further analysis revealed that there was a nucleotide sequence of 30 bp identical to the 30 bp sequence of the insert beyond which the event (insertion) occurred. The resultant repeated sequence concealed the chromatogram peaks generated by the Sanger sequencing and thus the Sanger's results were misinterpreted by us (Figure 3.7).




CLL-14, 7:82784832 C> G
VAF: 37.2%


CLL-5; FW
7:82784834-35 Ins30;
VAF: 23%


CLL-5; RV

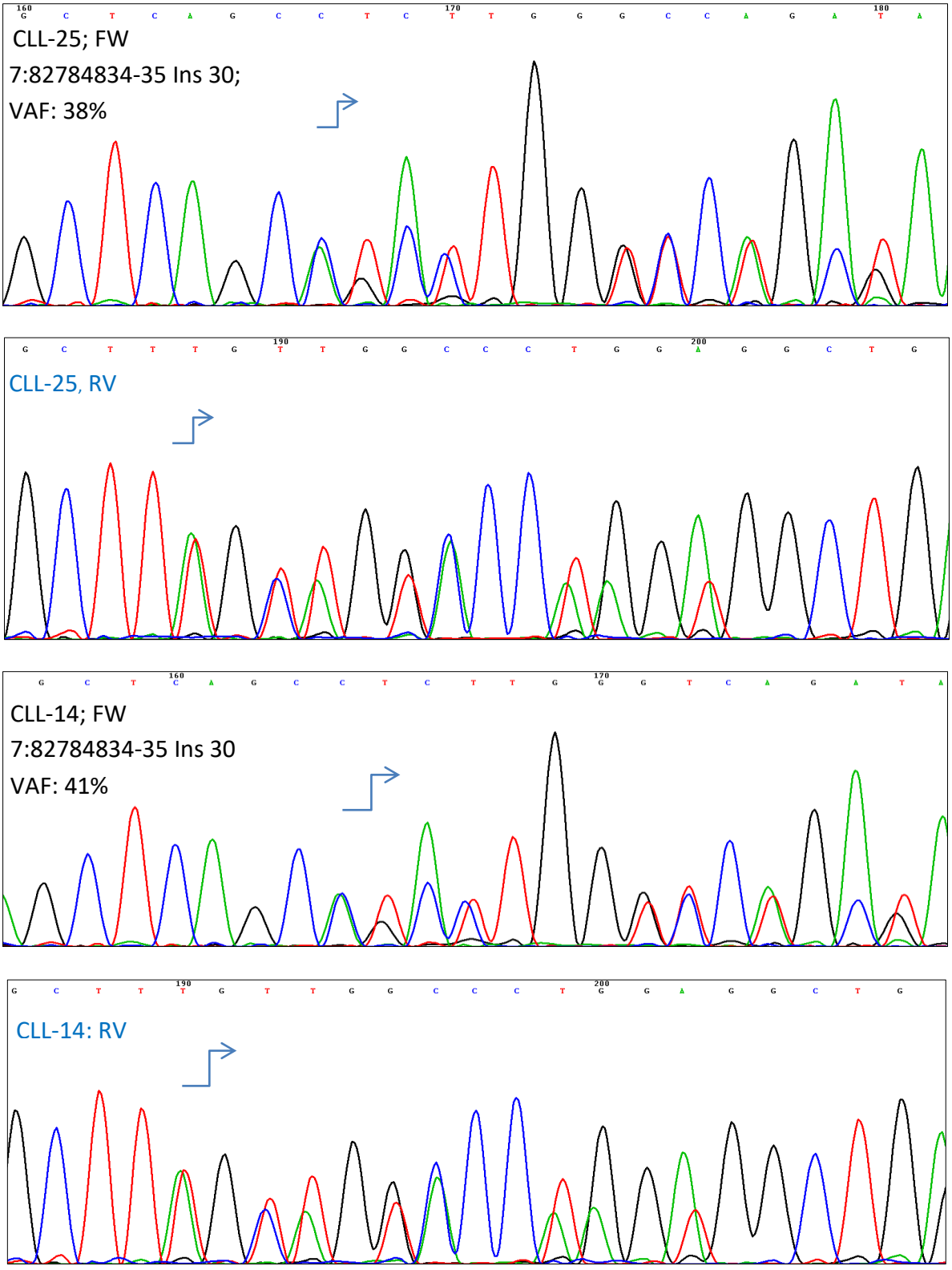
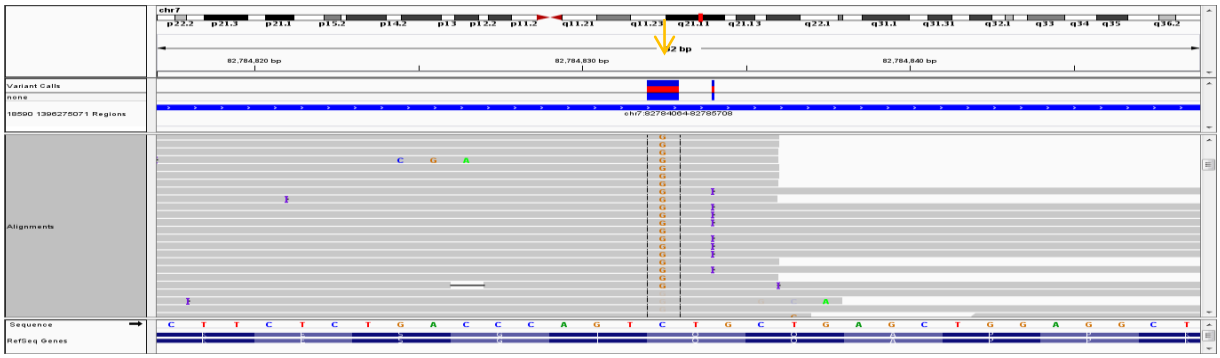


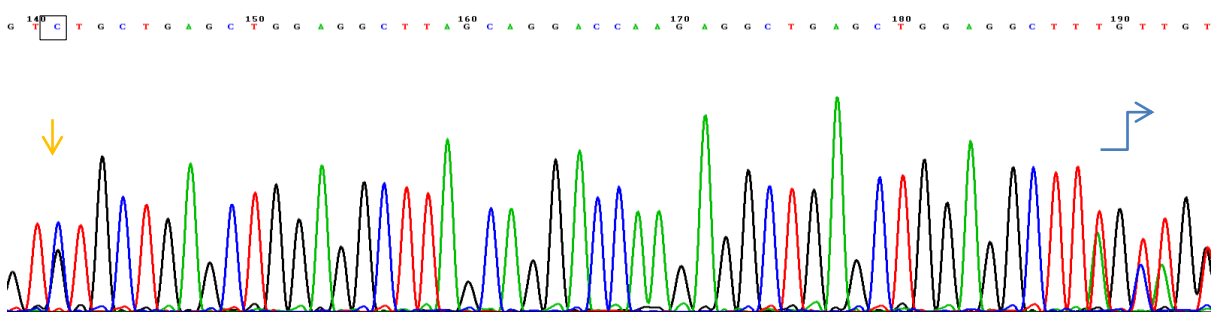
Figure 3.6. **Sanger sequencing for validation of 2 adjacent variants in *PCLO* detected repeatedly in 3 samples**

Highly consistent results for the point mutation were found, while the 30 bp insertion was consistent only in forward sequence and not in the reverse sequence.

A



B



C

The 30bp insertion sequence: CTCTTGGTCCTGCTAAGCCTCCAGCTCAGC

Forward sequencing >

Wild type (Wt.)allele(5'>3')
AACAAAGCCTCCAGCTCAGCCTCTTGGTCCTGCTAAGCCTCCAGCTCAGCAGACTGGGTCA
GAGAAGCCT

Mutant allele(5'>3')

AACAAAGCCTCCAGCTCAGCCTCTTGGTCCTGCTAAGCCTCCAGCTCAGCCTCTTGGTCCT
GCTAAGCCTCCAGCTCAGCAGACTGGGTCAGAGAAGCCT

Wt.allele AACAAAGCCTCCAGCTCAGCCTCTTGGTCCTGCTAAGCCTCCAGCTCAGC
AGACTGGGTCAGAGAAGCCT < Reverse
sequencing

Blue: Repeating sequence
Underlined blue: Insert sequence
Red: Mixed sequence

Figure 3.7. **Sanger sequencing and Ion Torrent PGM for a 30 nucleotides insertion in *PCLO***

A. The IGV snapshot of Ion Torrent PGM of a CLL case shows sequence reads (horizontal grey lines) carrying a point mutation C>G change gated in the middle of the window. The adjacent 30 bp insertion 2 nucleotides away from the right side of the point mutation appear as violent marks on the reads carrying the insertion. **B.** Sanger sequencing (reverse direction) chromatogram of the same case shows the point mutation C>G (marked by the box and the arrow in the left side of the window) and the 30 bp gap to the start of the insertion (marked by blue arrow in the right side of the window). **C.** The sequence of the insert, the Wt. allele and the mutant alleles are shown from 3'> 5'.

To sum up, so far, 13/14 (92.8%) and 10/12 (83.3%) of somatic non-synonymous mutations subjected to validation in respective replicated Ion Torrent experiments and Sanger based and /or As-PCR methods were confirmed as shown in Chapter 2 (Section 2.3.3.2). In addition, we have successfully confirmed 16/16 (100%) of somatic non-synonymous variants detected in other samples included in this chapter by PCR and Sanger sequencing. Accordingly, 39/42 (92.85%) of all variants subjected to validation were verified. That is, 100% of all variants (n = 24) with VAF% > 20% and 15/18 (83.3%) of those with VAF of < 20% (lowest detection limit of Sanger) were confirmed (Figure 3.8). As has been used in other studies [137], tracking of mutations in different samples from the same case supports mutations validity. Taking this in to account, additional to the above 39 validated mutations, sequencing of 23 serial samples (from the same patients but at earlier stages for study of clonal evolution in Chapter 4) tracked a total of 63/79 mutations. This strategy confirmed the existence of 32 extra mutations and extended the number of total validated mutations to 71/79 (89.8%).

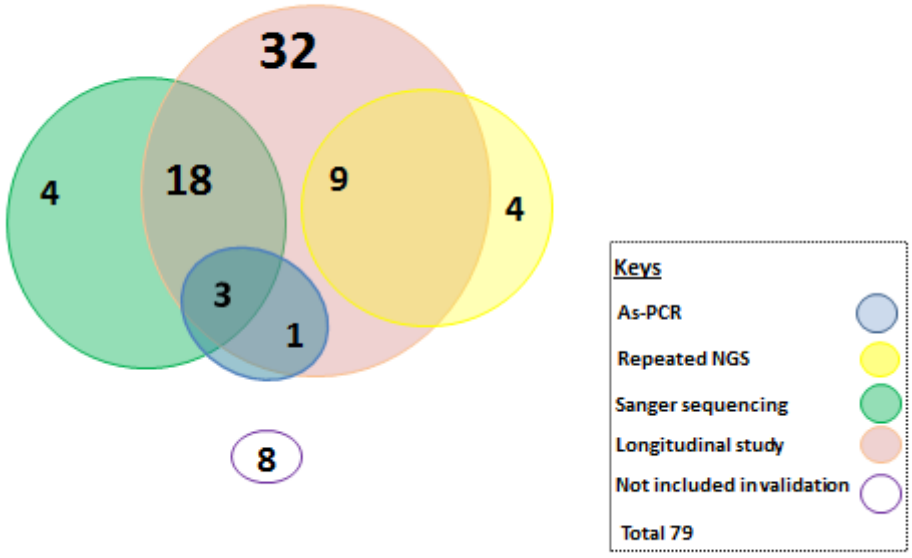


Figure 3.8. **Validated somatic non-synonymous mutations initially identified with the NGS method**

Venn diagram shows numbers of the mutations and methods used in the validation.

3.3.5. Somatic mutational profile in CLL with progressive and/or chemotherapy resistant disease

After the stringent variant calling and filtering (summarised in Figure 3.1), it was found that 87.5% (28/32) of the patients carried at least one somatic non-synonymous mutation in one of 12 targeted genes included in the 15 gene panel (Figure 3.11). Among the 12 targeted genes, 9 were mutated in more than one sample. Each of these mutated genes affected more than 5% of the patients in this cohort. Thus, *SF3B1* mutations occurred in 34.3% of the patients followed by *ATM* (31.2%) and *TP53* (28.1%). Other commonly mutated genes were *PCLO* and *NOTCH1*, which occurred in 25% and 15.6%, respectively. Each of *LRP1B* and *SAMHD1* mutations was found to occur in 9.3% of the cases, while each of *XPO1* and *FBXW7* in 6.2% of patients (Figure 3.9).

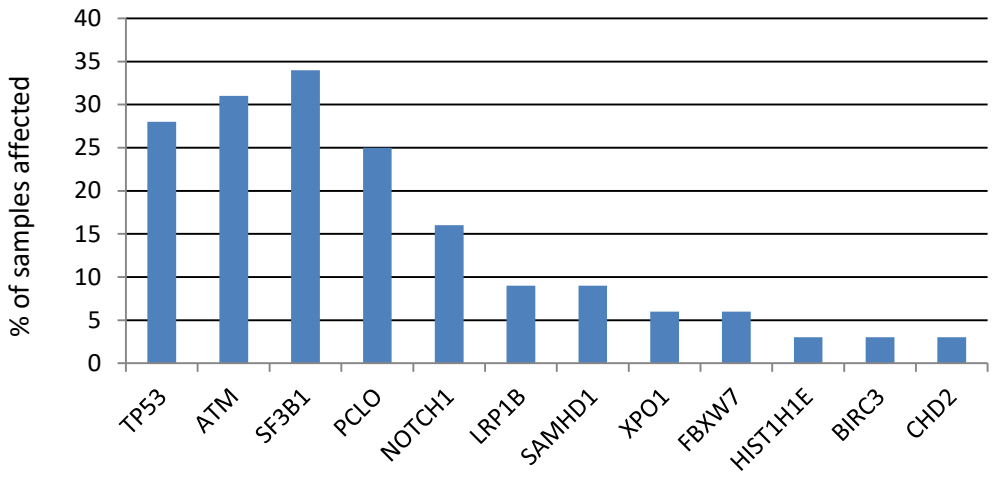


Figure 3.9. **Percentage distribution of samples bearing the mutated genes in the cohort of 32 CLL cases**

Regarding individual somatic non-synonymous mutations, 79.7% (63/79) of the mutations were unique for each sample and 20.2% (16/79) were recurrently detected in more than one case. The recurrent mutations occurred in 6 codons of 5 of the 12 genes analysed including *NOTCH1*, *SF3B1, ATM, PCLO,* and *XPO1*. Of the 28 patients harboured the mutations, 39.2% (11/28) had single gene mutations of which 36.3% (4/11) had at least another mutation in

the same gene. *TP53* and *ATM* mutations did not coexist in any samples. All patients with *TP53* mutations had at least another mutated gene, of whom, 66.6% (6/9) had mutations in *SF3B1* (Table 3.6). Moreover, 30% (3/10) of samples with mutated *ATM* had *SF3B1* mutations. That gave 81.82% of *SF3B1* mutations co-occurred with mutated *TP53* or *ATM*. Similar to mutual exclusiveness of *TP53* and *ATM* mutations, mutations in *NOTCH1* and its upstream gene *FBXW7* did not co-associate in any samples.

It is recognised that somatic mutation hotspots do exist in some genes, but not others. In agreement with this, our study found that 61.5% (8/13) of *SF3B1* mutations were localised in only 3 (exons 14 - 16) out of 14 exons sequenced and not surprisingly, 83.3% (5/6) of *NOTCH1* mutations affected a single codon p.F2482Ffs*2.

3.3.6. Somatic missense mutations predominated most of the targeted genes

In this cohort, various types of gene mutations were found including single nucleotide mutations and insertions or deletions of 1-30 nucleotides. Notably, missense single nucleotide variants were the most prevalent, accounting for 63.3% (50/79) (Figure 3.10). Missense point mutations were the most common mutations in *TP53*, while in *NOTCH1* and *BIRC3* no missense mutation was identified. Short indels prevailed *NOTCH1* mutations and the hotspot p.F2482Ffs*2 mutation was found in 5 of the CLL patients. Consistent with the finding by a previous study [273], indels were most prevalent among *BIRC3* mutations, we only found a single case affected by a single nucleotide deletion at codon p.V565fs*.

Considering *SF3B1*, 2 distinct nonsense mutations were identified; one of them was located outside the heat repeats (HR), while the other nonsense mutation located inside the HR2 and not the HR 5 - 8 which commonly occur in CLL. It has been found that mutations outside the HRs do not result in similar patterns of alternate splicing as found with mutations affecting the HRs [274]. Similar to findings reported by another study [275], *SF3B1* mutation p.K700E was the most recurrent mutation found in this cohort. Other hotspot mutations p.G742D (splice site variant) and p.K666M were found, but not repeatedly.

Table 3.6. **Somatic non-synonymous mutations detected in progressive and/or chemotherapy resistant CLL cases***

Genes	CLL cases																															Frequency	%
	CLL-1	CLL-2	CLL-3	CLL-4	CLL-5	CLL-6	CLL-7	CLL-8	CLL-9	CLL-10	CLL-11	CLL-12	CLL-13	CLL-14	CLL-15	CLL-16	CLL-17	CLL-18	CLL-19	CLL-20	CLL-21	CLL-22	CLL-23	CLL-24	CLL-25	CLL-26	CLL-27	CLL-28	CLL-29	CLL-30	CLL-31	CLL-32		
TP53	25	38	42	74	83	96	54	44	73																								9	28
ATM										50	40	98	10	7	18	42	48	50	5														10	31
SF3B1	8		4	50	47	53		2		6		30					2			80	41												11	34
PCLO	6	49			36			4						41										5	38	59							8	25
NOTCH1									81	21					8												12	24					5	16
LRP1B	6									6												48											3	9
SAMHD1														72					21				49										3	9
XPO1		25											24																				2	6
FBXW7							10												4														2	6
HIST1H1E																										45							1	3
BIRC3				13																													1	3
CHD2																			29														1	3
MYD88																																	0	0
ZFPM2																																	0	0
POT1																																	0	0
Total	4	3	2	3	3	2	2	3	2	4	1	2	2	3	2	1	2	1	4	1	1	1	1	1	1	2	1	1	0	0	0	0	56	

*Genes with multiple mutations in the same patient, only the mutation with highest allele frequency is presented. The cells left as blank grey if no mutation was identified

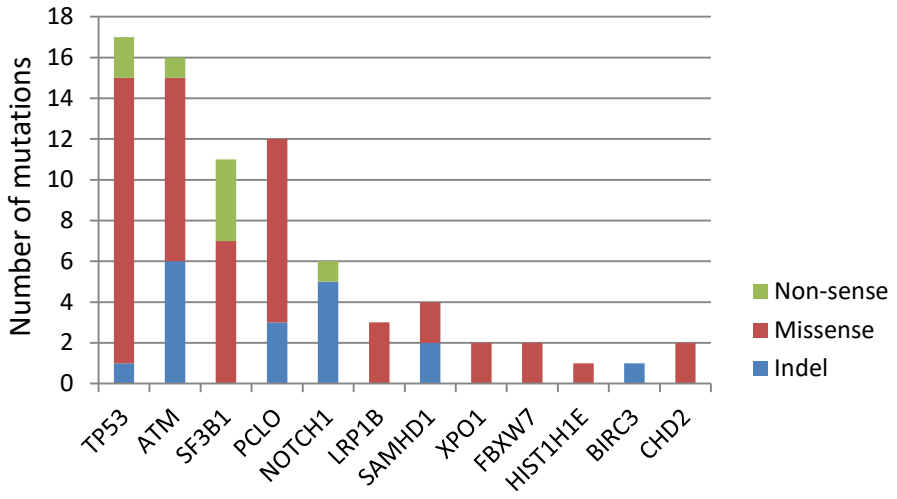


Figure 3.10. **Distribution of acquired somatic mutations by class across all genes analysed**

3.3.7. Frequency of mutant alleles

The variant allele frequency of 30.3% of the mutations (24/79) were below 10% and 37.97% (30/79) of the mutations were below 20% which is outside the detection limit offered by WES studies as well as Sanger sequencing, respectively. Consistent to findings by a previous study on a large cohort of CLL patients conducted by Jeromin et al [92], we identified 2 CLL samples that harboured *FBXW7* mutations with variant allele frequency below 10%. It is documented that even low levels of *FBXW7* gene mutation can exert oncogenic activity due to dominant negative effects which result in abnormal activation of NOTCH signalling (Figure 3.11).

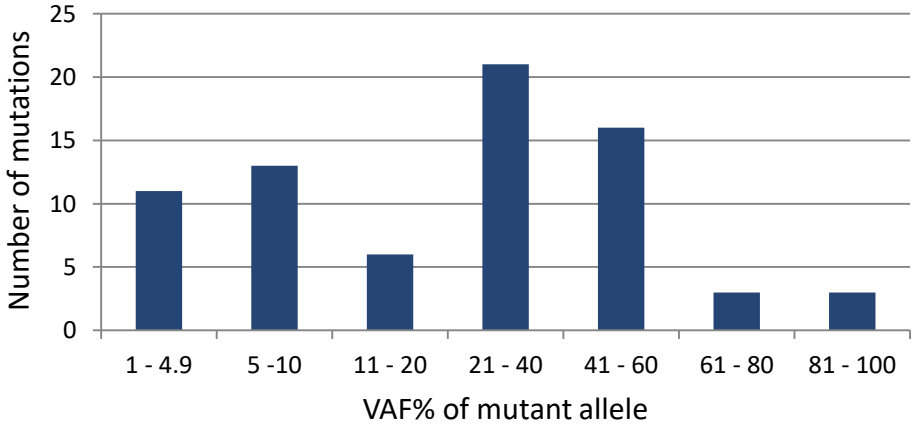


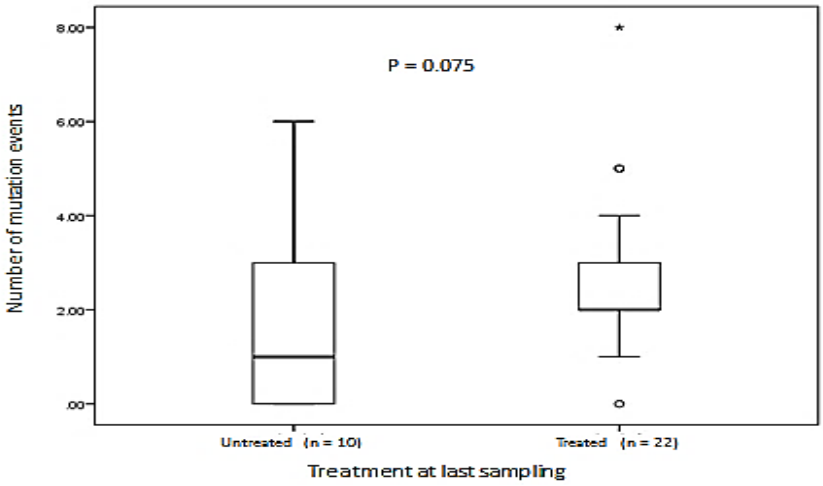
Figure 3.11. **Distribution of variant allele frequencies of the identified somatic non-synonymous mutations in targeted genes analysed**

3.3.8. Treatment history and its relation with the mutations

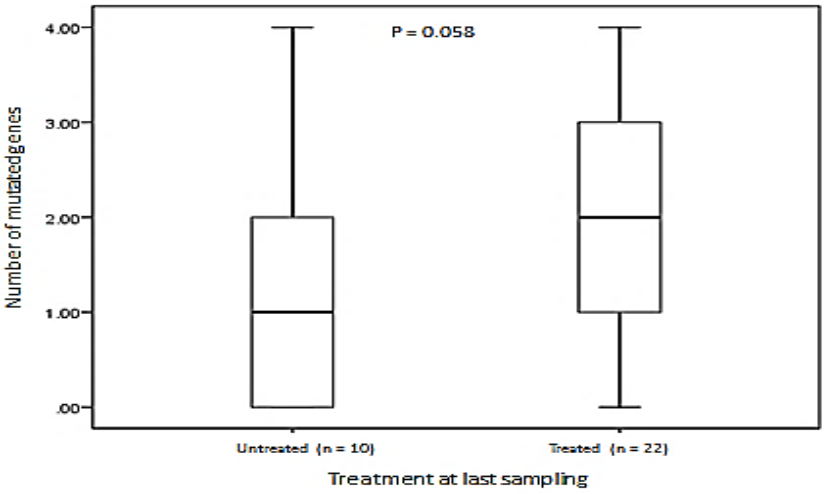
In this study cohort, 22/32 patients had received treatment with DNA damaging agents before the time of sampling, with mean and median number of treatment cycles being 3.48 and 4.0, respectively. To examine the impact of treatment on mutation, we firstly compared the number of mutation events (Figure 3.12.A) and the number of mutated genes (Figure 3.12.B) identified among the cases at the latest stages between cases with and without the previous chemotherapy. The untreated group (n = 10) had a lower number of mutation events median: 1 (range: 0 - 6) compared to the treated group (n = 22), median: 2.0 (range: 0 - 8), although the difference was not statistically significant, P = 0.075 (Mann-Whitney test). Likewise, we examined the number of mutated genes between the untreated and the treated group, the former had a lower number of mutated genes median: 1.0 (range: 0 - 4) compared to the latter group median: 2 (range: 0 - 4), with the P value (0.058, Mann-Whitney test) being close to the α level.

Next, we explored the interrelationship between the number of treatment cycles and the number of mutated genes. We applied the median number of treatment cycles received at time of sampling in this cohort as a cut-off to divide the cohort into two groups. Hence, the number of patients with ≤ 3 cycles of chemotherapy was 15 and that with > 3 cycles was 17. When counted cases with $\geq$ 2 mutated genes (the median in this cohort), we found a statistically significant difference between the 2 groups (P = 0.036, χ^2 test) (Figure 3.12.C). Thus, more than one mutated genes were identified in 13/17 (76.5%) cases who received > 3 cycles of therapies, but only in 5/15 (33.3%) of the remaining patients.

A



B



C

	Chemotherapy		P
	≤ 3 cycles	> 3 cycles	
Samples with < 2 mutated genes	10	4	0.036
Samples with ≥ 2 mutated genes	5	13	0.036

Figure 3.12. **Relationship of chemotherapy with somatic mutations**

Comparison of the number of mutation events **A.** and number of mutated genes **B.** between patients with and without previous chemotherapies. Comparison of the frequency of multiple mutated genes between cases received at least 3 chemotherapy cycles and the remaining cases **C**.

3.3.9. Gene mutations in ATM/p53 and other pathways

When we analysed somatic mutations in signalling pathways, we found that those in genes involved in RNA processing and splicing, *SF3B1* and *XPO1*, occurred in 13 patients, but they were dominant, being either the only mutated gene or the one with the biggest VAF, in only 3 (23.1%) samples. Mutated genes involved in the NOTCH signalling pathway, *FBXW7* and *NOTCH*1, were identified in 2 and 5 samples, respectively. Only, *NOTCH1* mutations were dominant in 3 (42.9%) samples. Mutations in histone modification genes, *CHD2* and *HIST1H1E* occurred in 2 samples. Only the *CHD2* mutation predominated in the multigene mutated sample. Mutations in NFκB signalling genes *BIRC3* and *SAMHD1* occurred in 1 and 3 patients, respectively. *SAMHD1* mutations were dominant in 2 samples. However, the tumour suppressor and DNA repair gene *TP53* and its upstream activator *ATM* were the most commonly mutated genes detected in 59.3% (19/32) of the sample cohort (Table 3.6). More importantly, both of them were the dominant mutant genes detected in 14 of the 19 (73.7%) samples.

Chi-Square test confirmed that the dominant mutations in *ATM* and *TP53* were significantly more frequent than that in other genes studied in this chapter (73.7% versus 37.8%, P = 0.011) (Figure 3.13). These findings suggest defective DNA repair pathway play a dominant role in genomic instability and may facilitate acquisition of additional mutations in CLL.

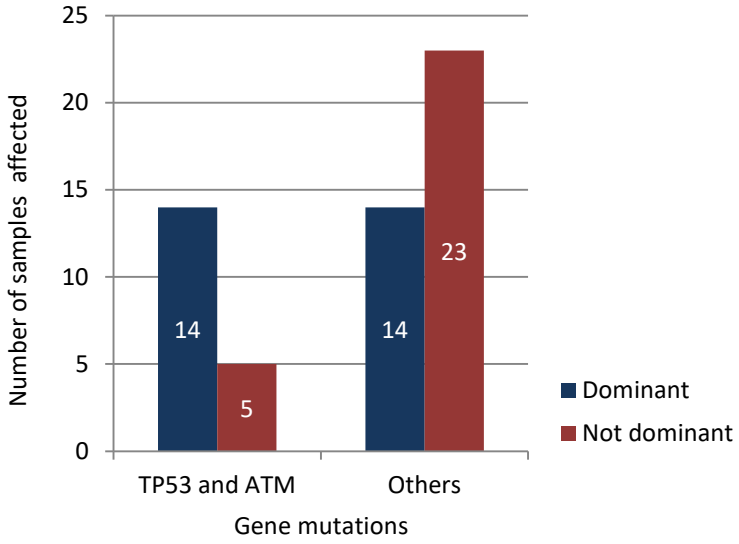


Figure 3.13. **Comparison of dominant somatic mutations in *ATM* and *TP53* and other genes in samples from patients with advanced CLL**

The dominant mutation was defined as that was either the only mutated or the one with the biggest VAF.

3.3.10. SNP array analysis showed high quality of data of the whole genome copy number changes

As mentioned in Section 3.2.8, 14 mutated CLL samples from the cohort were subjected to genome-wide SNP array analysis. CytoSNP-850K array data quality control was measured using the data quality metrics. Visual inspection of Log R and B-allele frequency charts of sample CLL-8 revealed high levels of genomic background noise (Figure 3.14). This sample was therefore excluded from subsequent analysis. Data from all of other samples showed good quality similar to sample CLL-7 as shown in Figure 3.14.

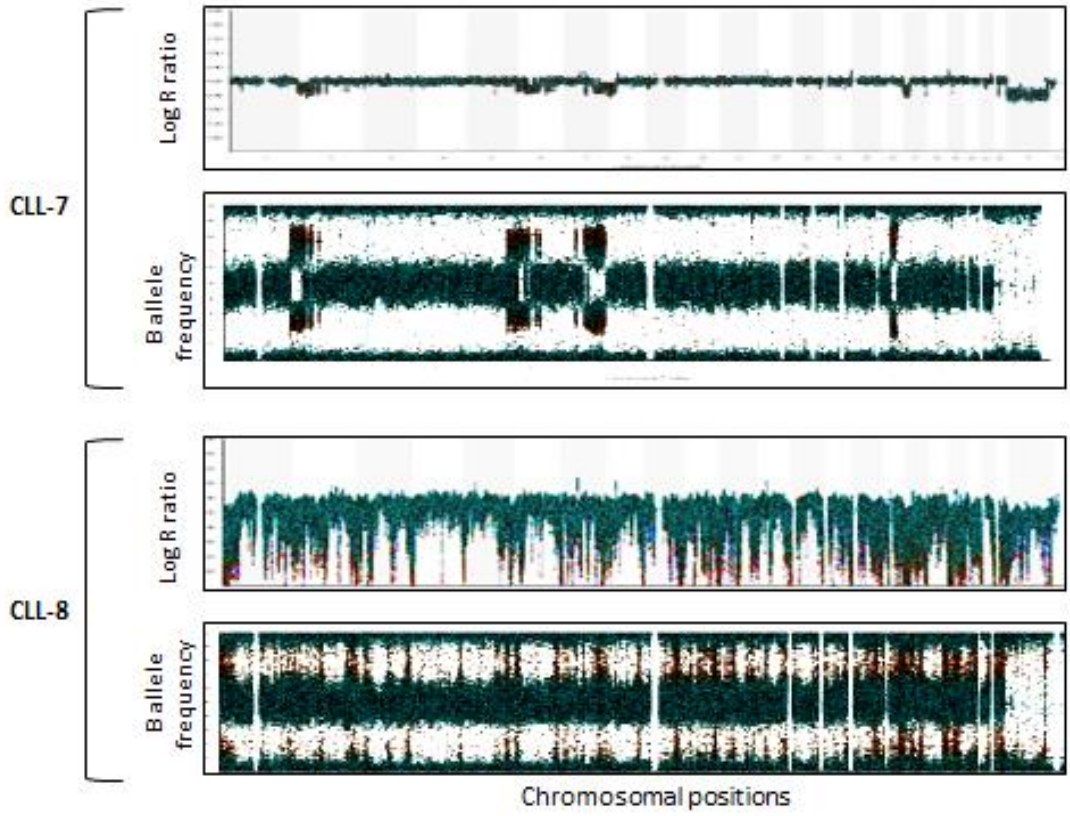


Figure 3.14. **Examples of good and poor quality SNP array data from visual display of Log R ratio and B allele frequency**

Log R ratio chart of CLL-7 shows acceptable noise and clear negative shifts from the baseline (close to 0) in some regions indicating large-scale deletions. Accordingly, the B allele frequency chart of (CLL-7) show clear deviations from the baseline in the same regions. For CLL-8, high levels of genomic background noise resulted in loss of clear cut demarcation of normal and abnormal genomic status.

Genotyping accuracy (as estimated by calculation of the frequency of heterozygous SNP calls of male X-chromosome SNPs) was 100%, P = 0.001. Furthermore, all the 8 recurrent copy number aberrations previously identified by FISH were also correctly profiled in the corresponding samples, as shown in Table 3.7.

Table 3.7. **An overview of X chromosome and previous FISH analysis used to assess sensitivity of the CytoSNP-850K BeadChip array**

CLL cases	Gender	Identified No. of X-chromosome by CytoSNP-850K array	FISH results available	Identified by CytoSNP-850K array
CLL-1	male	1		
CLL-3	female	2	Biallelic 13q14.2-	yes
CLL-4	male	1		
CLL-7	male	1	Monoallelic 17p13.3-	yes
CLL-10	female	2		
CLL-11	female	2	Biallelic 13q14.2-	yes
CLL-12	male	1	Monoallelic 13q14.2- and 11q21.23-	yes
CLL-17	male	1	Monoallelic 13q14.2-	yes
CLL-19	male	1	12+	yes
CLL-20	male	1	12+	yes
CLL-21	female	2		
CLL-23	male	1		
CLL-26	Female	2		

3.3.11. SNP array analysis identified recurrent and non-recurrent copy number aberrations in CLL

Chromosomal abnormalities were detected in 92.3% (12/13) of the patients. 10 of them had at least 2 aberrations and 7 had complex karyotypes (as defined in Chapter1, Section 1.2.3.3). Chromosomal gains (n = 9) were less frequent than chromosomal losses (n = 34). As shown in Figure 3.15, among the recurrent copy number changes in CLL, 13q114.2 - q14.3 deletion was the most common cytogenetic abnormality and occurred in 6 patients (monoallelic loss in CLL-1, CLL-4, CLL-12 and CLL-17 and biallelic losses in CLL-3 and CLL-11). This lesion did not occur as solitary abnormality in any cases. The median size of this deletion was 3 Mbp although only in one CLL sample a large deletion spanning 27 Mbp was identified that affected the *RB1* locus (CLL-4).

Moreover, trisomy 12 was the 2nd most common cytogenetic abnormality in this cohort occurred in 4 patients, 75% of which occurred in combination with other cytogenetic changes. 17p13.3 - p11.2 deletion occurred in 2 cases harbouring *TP53* mutations. In addition, a smaller region of 9,150,800 bp was affected by copy neutral loss of heterozygosity (CNN-LOH) in another patient with *TP53* mutations. Sample CLL-1 harboured *TP53* mutations with VAF below 25%, but did not have detectable 17p abnormality. Hence,

3/4 (75%) of the patients with *TP53* mutations lost the wild-type allele and all of these three cases harboured complex karyotype.

Considering 11q abnormalities, loss of 11q21 - q23.3 occurred in 2 patients (CLL-4 and CLL-12), who carried *BIRC3* and *ATM* mutations, respectively. Conversely, CNN-LOH of 11q13.2 - q25 and gain of 11q22.3 - q25 occurred in another 2 patients LL-26 and CLL-7, respectively, who lacked identifiable mutations in the two targeted genes mapped on 11q.

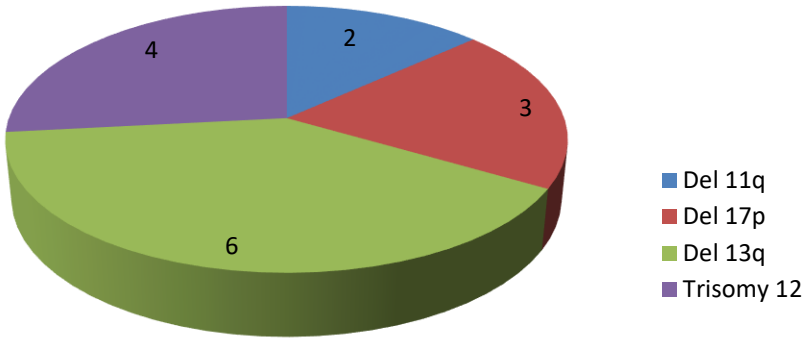


Figure 3.15. **The types and numbers of recurrent cytogenetic aberrations detected with the SNP array assay in the 13 CLL samples**

Notably, 69.2% (9/13) of the patients had at least one genetic aberration other than the recurrent copy number abnormalities mentioned above (Table 3.8). Interestingly, some of these affected chromosomal arms were identified in multiple samples, with the same or different region sizes. For instance, loss of 6p25.2 - p25.3 spanning from 3,077,141 - 4,394,814 which harbour interferon regulatory factor *IRF4* locus occurred in 2 patients (CLL-7 and CLL-17). However, losses of 6q mapped between 6q11 and 6q27 or between 6q14 and 6q15 were identified in 3 patients. Similarly, deletions of 7q21.11 -7q34 and 7q33 - 7q34

were detected cases CLL-7 and CLL-26, respectively. Additionally, a gain of another copy spanning the entire length of chromosome 7 was identified in CLL-4.

Losses of chromosome 8 short arm (8p12 - 23.3) were identified in CLL-7 and CLL-11. Gains of 8q23.3 - q24.3 and 8q21.11 - q24.3 were found in CLL4 and CLL-11, respectively. CNN-LOH of 9q21.11 - q34.3 (carrying *NOTCH1* locus) occurred in one sample (CLL-10) which harboured *NOTCH1* mutations. Similarly, CNN-LOH of 2q14.3 - q37.3 (containing *SF3B1* locus) occurred in one patient with *SF3B1* mutations. All the other non-recurrent chromosomal aberrations did not structurally associate with somatic mutations detected in this study. A possible chromothripsis event affected chromosome 6 (5 oscillating loss of heterozygosity affected the 6p) (described in Chapter 1, Section 1.2.3.5) and occurred in one patient who harboured *TP53* mutations and 17p 13 deletions (CLL-7). However, for definitive chromothripsis we need to identify that the event occurred as a single cellular catastrophe on the same homolog and not as multistep accumulation of cytogenetic changes on different homologs [276]. Details of each chromosomal aberration start and end points are presented in Appendix 7.3.6.

Table 3.8. **Integrated targeted NGS and high resolution genome wide SNP array analysis***

Genes	CLL-1	CLL-3	CLL-4	CLL-7	CLL-10	CLL-11	CLL-12	CLL-17	CLL-19	CLL-20	CLL_21	CLL-23	CLL-26
TP53	25	42	74	54									
ATM					50	40	98	48	5				
SF3B1	8	4	50		5		30	2		80	41		
PCLO	6												59
NOTCH1					21								
LRP1B	6				6								
SAMHD1									21			49	
FBXW7				10					4				
HIST1H1E													45
BIRC3			13										
CHD2									29				
Chromosomes													
2p				2p25.3-11.2									
2q										2q14.3-37.3			
4p			4p16.3-15.1										
4q	4q21.1-21.3		4q34.2-35.2										
6p				6p22.3-11.2				6p25.3-25.1					
6q				6q13-15				6q14.1-27		6q14.1-21			
7p			7p22.3-										
7q			7p22.3-	7q31.1-36.3									7q33-34
8p				8p23.3-12		8p23.3-12							
8q			8p21.1-24.3			8p21.1-24.3							
9q					9q21.11-34.3								
10q				10q24.1-26.3									
11q			11q22.1-24.1	11q22.3-25			11q21-23.3						11q13.2-25
12p					12p13.33-				12p13.33-	12p13.33-		12p13.2-13.1	12p13.33-
12q					12q24.33				12q24.33	12q24.33			12q24.33
13q	13q14.2-14.3	13q14.2-14.3	13q14.2-22.2			13q14.2-14.3	13q14.2-14.3	13q14.2-14.3					
17p		17p13.3-11.2	17p13.3-13.1	17p13.3-11.2									
18p								18p11.32-11.21					
18q		18q22.1-23	18q11.2-23										
19p				19p13.11-13.3									
20q										20q11.21-13.33			

Keys for CNA
Gain
CNN-LOH
Biallelic Loss
Monoallelic Loss
Chromothripsis-like event

* The displayed number for each gene is VAF% of the dominant mutation identified if multiple mutations existed in a sample and colour coded by the CNA identified. Only the chromosome aberration with the biggest size in one chromosome arm is presented for each sample.

The novel (non-recurrent chromosomal aberrations) chromosomal aberrations were also analysed for their genomic contents, a large proportion of cancer genes and some CLL mutated genes listed in Chapter 1 (Section 1.2.4.3) were found to be affected. Notably, a large proportion of DNA repair and cell cycle control genes were affected as shown in Table 3.9.

Table 3.9. **Novel CNAs identified in the CLL cohort with SNP array analysis**

Cytogenetic aberration	**Biggest affected band identified by the SNP array**	**Genes**	**Function** (NCBI,2016)
2p-	2p11.2 - 2p25.3	*ADAM17* *MYCN* *EFEMP1* *HTRA*	Cell cycle control Cell cycle control Cell cycle control Cell cycle control
2q-	2q14.3 - 2q37.3	*CXCR4* *IL1B* *ITGA6*	Lymphocyte differentiation Lymphocyte activation Cell migration
4P-	4p15.14 - p16.3	*CD38*	Apoptosis
4q-	4q21.1 - 4q35.2	*FAT1*	Control cell proliferation
6p-	6p11.2 - 6p25.2	*IRF4* *RIPK1* *NFKBIE* *PIM1*	Transcription activation Inflammatory and apoptotic pathway B cell inhibitor Proto-oncogene in B cell lymphoma
6q-	6q14.1 - 6q27	*NT5E* *TNFAIP3*	Lymphocyte differentiation
7q-	7q21.11 - 7q36.3	*HGF* *ING3* *BRAF* *CAV1* *EZH2*	Cell cycle control Cell cycle control Cell cycle control Cell cycle control Transcription control
8p-	8p12 - *8p23.3*	*WRN* *MTUS1*	DNA repair Tumour suppressor
8q+	8q21.11 - 8q24.3	*ZFPM2* *MYC* *NDRG1*	Transcription control Cell cycle control Cell cycle control
9q CNN-LOH	9q21.11 - 9q34.3	*SMARCA2* *TLR4* *NOTCH1*	Chromatin regulation Immune regulatory pathway NOTCH pathway
20q CNN-LOH	20q11.21 - 20q13.33	*ASXL1*	Transcription regulator

3.3.12. Relationship of CNA with gene mutations and chemotherapy

With SNP array data of the 13 patients available, we then tried to examine any possible impact of previous chemotherapy on the combined number of mutation events and the CNA. There was no statistically significant difference between the two groups. This was most likely due to the small size of samples and the bias of distribution of patients in both groups to be compared. In fact, only one case in this cohort was untreated at time of sampling and only 2 cases had received treatment ≤ cycles at time of sampling.

Moreover, there was no statistical correlation between the number of CNA and the number of mutated genes detected. However, the combined number of somatic mutation events and the CNA was significantly different (P = 0.026, Mann-Whitney test) between samples with *TP53* or *ATM* mutations, median: 8 (range: 4 - 19) and those with other gene mutations median: 4 (range: 1 - 5) (Figure 3.16).

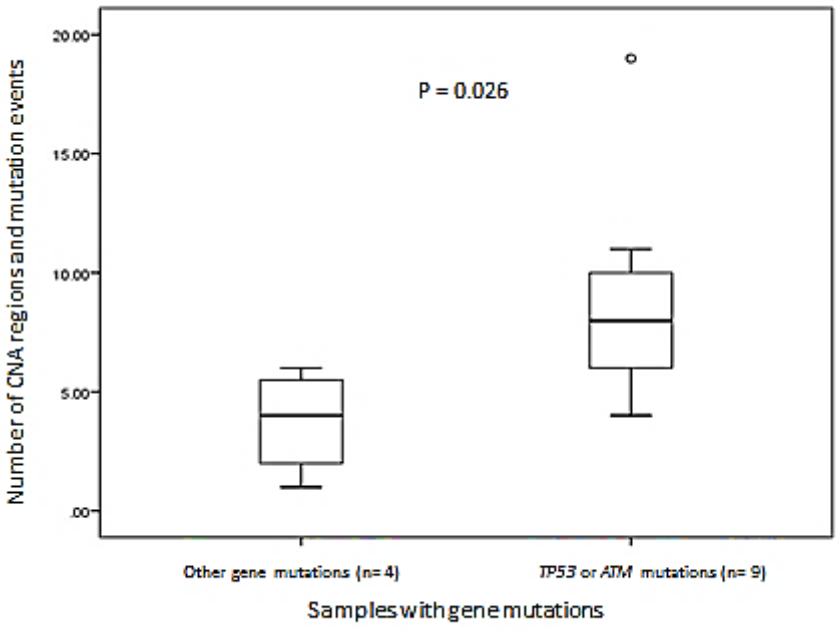


Figure 3.16. **Relationship of combined number of CNAs and target gene mutation events with *TP53 or ATM* mutations in the cohort of 13 CLL patients**

3.4. Discussion and conclusions

At the beginning of this work, two earlier high throughput studies identified novel recurrent gene mutations and CNAs in a limited number of unselected CLL patients [245, 277]. It was unclear how these somatic mutations and CNAs are distributed across CLL patients with aggressive clinical phenotype. As exploratory studies, they were designed to globally screen genomic aberrations with WGS or WES, and therefore did not provide information about subclonal architecture and its relations with disease progression and relapse of CLL. Moreover, information on copy number alterations from these studies was limited as they were identified with FISH or low resolution SNP array. In our project, we designed a longitudinal study of progressive and/or chemo-resistant CLL using an ultra-deep targeted NGS technique and a more sensitive SNP array assay to deepen our understanding of somatic mutation pattern, subclonal architecture and their interrelations throughout the disease course. In this chapter, we started our study with establishing a profile of these recurrent genomic aberrations in a cohort of patients with progressive and/or chemo-resistant CLL. This was necessary for our further study on mutated subclonal evolution in Chapter 4.

Firstly, our analysis of coverage depth, the stringent quality control assessment and the successful validation of candidate variants with additional methods suggested the high and reliable quality of sequencing data produced by the NGS method. Although our gene panel did not include all recurrent mutated genes identified subsequently by other centres [272, 278-280], the somatic mutation profile of the 32 CLL patients with advanced CLL clearly showed that 87.5% of them harboured at least one mutation in 12 of the 15 selected genes. This indicates the importance of these recurrent mutations and their involvement in pathogenesis of the disease. As shown in Figure 3.11, 30.3% of the detected mutations had variant allele frequency below 10% which is below the detection limit of WGS or WES and Sanger sequencing. This emphasises the importance of deep NGS techniques, so that the variants with low allele frequency can be identified at an early stage of the disease. Moreover, *TP53* and *ATM* mutations dispersion throughout the coding sequences explains the unfeasibility of Sanger sequencing for screening either large genes or multiple genes in a single test and further underscores the necessity of large-scale analysis obtainable through NGS techniques.

Consistent with the findings in a recent study in CLL with targeted deep sequencing (published in 2015) [231], our study showed that missense single nucleotide variants predominated in the somatic mutations of the majority of the genes, except *NOTCH1* and *BIRC3* in which indels were most prevalent. In agreement with L. Wang's finding in 2013, we found *SF3B1* as the most commonly mutated gene in patients with progressive and therapy resistant CLL [281]. However, most of them co-existed with mutations in other genes, especially *TP53 and ATM* (see below). However, in the current study two distinct nonsense mutations were identified with a variant allele frequency below 10%, of which one was located outside the Heat Repeat region, emphasising a necessity to extend the target region beyond the hot spot in studies with deep sequencing approaches. Not surprisingly, no somatic mutations in *MYD88* were detected in this cohort of samples. This was possibly because all samples tested by us had been taken at a stage of disease progression and/or treatment resistance and because the *MYD88* mutations predominately occur in patients with favourable prognostic features [169, 282].

The most interesting finding in this chapter was the possible role of mutations in two genes involved in cell-cycle and DNA-repair pathways namely, *ATM* and *TP53*, and their association with other genetic alterations. Firstly, most mutations in *ATM* and *TP53* have been reported in patients with aggressive diffuse large B cell lymphoma as recorded in the COSMIC data base, indicating their driving roles in CLL progression. Secondly, they were among the most common somatic mutations identified in this study cohort, accounting for 41.77% of the total mutation events (Table 3.6). Thirdly, in most cases they were dominant, in terms of clone size over co-exiting mutations in other genes, including the most commonly affected *SF3B1* and *PCLO* (Section 3.3.9 and Table 3.6). Fourthly, samples with *ATM* or *TP53* mutations harboured more genomic aberrations, including somatic mutations detected with the NGS method and CNA as detected with the SNP array, than other samples (Figure 3.16). Furthermore, a clear trend of positive correlation between previous chemotherapy and the number of somatic mutations was found in this cohort of patients in different statistical analyses (Section 3.3.8), although some P values were just above the pre-set α level (0.05) possibly due to the small size and/or biased distribution of samples. Giving the well-known functions of the ATM-p53 pathway in DNA repair and genomic stability, our results strongly suggest that defects in this pathway contribute to acquired genomic aberrations and

therefore the clonal evolution. Whether the clonal evolution resulted from new mutations induced or pre-existing mutations selected by the DNA damaging therapies will be addressed in the longitudinal study in Chapter 4.

Using the high resolution SNP array assay, we were able to detect both the commonly recurrent (13q-, 11q-, 17p- and 12+) and novel chromosome CNAs in CLL. As expected, we found 13q deletions to be the most common abnormality in this cohort of 13 patients studied. All the 13 q deleted samples had at least another chromosomal aberration; only in one sample (CLL-4) the *RB1* locus was affected. Alteration of this gene locus is found to be associated with fast disease progression in CLL [283]. In two other patients biallelic 13q14 deletions were found. However, the significance of clinical impacts of biallelic 13q deletion remained controversial [284], and we could not exclude any role of the co-existing gene mutations and/or other CNAs. In addition, novel cytogenetic changes were identified. They affected chromosomes 2, 4, 6, 7, 8, 9, 10, 18, 19 and 20. Some of them were found in more than one sample (Table 3.8), suggesting possible hotspots of CNAs in CLL patients. In these regions, genes related to human cancers including CLL, were identified (Table 3.9). They play role in cell cycle control, DNA repair, chromatin modification, inflammatory and cell migration. These findings emphasize the roles of these pathways in CLL pathogenesis [177, 185, 285].

The heterozygous deletion of 17p, 2q and 9q resulted in CNN-LOH of *TP53*, *SF3B1* and *NOTCH1* in patients with mutations in those genes, respectively. Although this has been previously reported for *TP53* and *SF3B1* mutations, there has not been any report for *NOTCH1* mutations in CLL. Despite of no additional predicting value of LOH for *TP53* in CLL, it is still not clear if the combined deletion and mutation will accelerate CLL progression in patients with LOH of *SF3B1* or *NOTCH1.*

Regarding the clinical impact of the gene mutations found in this cohort of CLL samples, our statistical analysis did not show any significant difference in treatment free survival (TFS) calculated from diagnosis to treatment or death and overall survival (OS) calculated from diagnosis or time of test to the last follow-up or death between patients with different gene mutations (data not shown). This might be because the mutation profile presented at the latest stage of the disease has no predictive value for the TFS and the OS calculated from the

time of diagnosis. It might be also not suitable for the OS calculated from time of sampling, as most of those case died shortly after the last sampling. We therefore planned to correlate clinical outcome of those patients with mutation profiles found at early stages in the longitudinal study in the Chapter 4.

Chapter 4. A longitudinal study of mutant clonal evolution using targeted ultra-deep NGS in patients with progressive and/or chemo-resistant CLL

4.1. Introduction and aims

As mentioned earlier, the WGS studies have revealed a number of recurrently mutated genes in CLL shortly before the start of this study. Subsequent studies with combined chromosomal copy number change [242, 286] and high throughput NGS sequencing techniques [137, 277] have suggested that CLL cells in a patient can bear different subclonal mutations or copy number changes that may evolve over disease course. More recently, the application of highly sensitive NGS allowed detection of a small size of *TP53*-mutated subclones and its clinical relevance has been elucidated [138]. In addition, the presence of subclonal driver mutations in other genes has been suggested as an influential factor for a clinically aggressive CLL phenotype [198].

Despite a growing body of evidence for the association of genomic alterations and CLL clinical heterogeneity, the role of these minor mutated subclones involved in other recurrently mutated genes and their evolution dynamics are still not well understood. This might be partly because small subclonal mutations are underestimated due to limited sensitivities of the employed techniques. However, these subclonal mutations possibly exist at earlier stages or before disease progression and therapy resistance in CLL.

Therefore, for a better understanding of the order of gain and selection of these genomic aberrations, we applied a more sensitive deep sequencing technique in this chapter. Primarily, we planned to reliably explore subclonal diversity and define the evolution pattern of these subclones over the period of disease progression and/or relapse. Secondly, we aimed to explore the possible mechanism of mutation operating in CLL evolution and relapse by identifying the pattern of somatic mutations before and after therapy. We analysed sequencing data of serial samples from the 23 mutated CLL patients with aggressive clinical

phenotype to address these questions. Such information might be helpful in developing novel biomarkers for better stratifying patients (predicting disease progression and resistance to treatment) at early stages. In addition, we applied genome wide high resolution SNP array for integrated analysis of evolution at chromosomal levels in a CLL patient with 2 *TP53* mutations that differentially affected p53 function.

4.2. Materials and methods

4.2.1. CLL samples

At the time of study for this chapter, samples from 25 of the 28 mutated cases (in Chapter 3) taken at earlier stages were selected. We selected 38 such samples from these 25 cases for this longitudinal study. These samples covered different time-points (range 2 - 4 time points) of the disease course including stable phase, disease progression prior to chemotherapy, and relapse. They were cryopreserved either as isolated CLL cells (inside the Bio-bank) or as prepared g. DNA (outside the Bio-bank). The clinical information for these samples is summarised in Table 4.1.

Moreover, the 2 DNA samples from CLL-7 taken at time of diagnosis and after disease progression but before chemotherapy administration were also studied for integrated analysis of copy number changes and gene mutation evolution over disease progression and after relapse.

Table 4.1. **Clinical information of the additional time points of 23 CLL cases analysed for clonal evolution***

CLL cases	Time point 1			Time point 2			Time point 3		
	Lymphocyte $10^3/\mu l$	Binet stage	Phase/time of sampling	Lymphocyte $10^3/\mu l$	Stage	Phase/time of sampling	Lymphocyte $10^3/\mu l$	Stage	Phase/time of sampling
CLL-1	160	B	Relapse						
CLL-2	108	B	Stable after Dx						
CLL-3	116.9	A	Stable after Dx	81	NA	Progression			
CLL-4	39.5	NA	Dx						
CLL-5	NA	A	Progression						
CLL-6	23.8	C	Start of 1st Tx						
CLL-7	7	B	Dx	61.6	B	Start of 1st Tx			
CLL-8	43.7	A	Dx	129.9	C	Relapse			
CLL-10	9.7	A	Dx						
CLL-11	120	A	Dx						
CLL-12	90	C	Dx and start of Tx	104.1	C	Relapse			
CLL-13	127	A	Remission	316	C	Relapse			
CLL-14	147	B	Dx						
CLL-15	12.7	A	Stable after Dx						
CLL-17	184	B	Stable after Dx	10.3	B	Remission	263	C	Relapse
CLL-18	13.6	B	Dx						
CLL-19	36.7	B	Remission						
CLL-20	123.3	B	Dx						
CLL-21	120.1	A	Dx	287.4	B	Start of 1st Tx			
CLL-22	129.8	A	Dx	268	C	Start of 1st Tx			
CLL-23	314	A	Stable after Dx						
CLL-25	102	A	Dx						
CLL-26	35	A	Dx	110.4	B	Start of 1st Tx			

*Dx: diagnosis; Tx: chemotherapy, NA: information not available. Stable after Dx: If the sample was not taken at date of diagnosis, but no progression occurred until the date of sampling. Grey filled spaces: no further serial time points were sequenced.

4.2.2. DNA preparation, target enrichment, deep sequencing and data analysis

These procedures were the same as described in Chapter 2, Materials and Methods Section 2.2.

4.2.3. CytoSNP array for monitoring of chromosomal copy number changes

In collaboration with Merseyside and Cheshire Regional Genetic Laboratory, the 2 DNA samples mentioned in section 4.1, were subjected to CytoSNP 850K array (Illumia, UK) as described in Chapter 3, Section 3.2.8.

4.2.4. Statistical analysis

Statistical analysis was conducted using IBM SPSS v21. Medians and ranges were presented for variables with skewed distribution. Statistical significance of association with clonal evolution between AID-, ageing- and other factor-related mutations were assessed using Chi-square test. Wilcoxon- Rank test was used for testing significance of changes occurred in follow up for each group of mutations. ANOVA was applied to test the significance of difference in mutation doubling time (from the earliest to latest time of sampling) between multiple gene mutation groups. Mann-Whitney test was used to test the significance of difference between two independent groups. For testing clinical impact of gene mutations, Kaplan-Meier survival curves and Log-rank test were used. Overall survival (OS) was defined from the diagnosis or time of first sampling to the time of death (as the event) or last follow-up (censored). Treatment free survival (TFS) was measured from the diagnosis or time of first sampling to the time of first treatment/death (as the event) or last follow-up (censored). Statistical significance was defined as $p < 0.05$. All P values were calculated in double sided test.

4.3. Results

4.3.1. SNP fingerprints intrinsically controlled tracking of serial samples

Before tracking the mutant CLL clones in the serial samples sequenced, the samples had to pass a number of quality filters to ensure reliable data. These filters included all the sequencing quality control metrics (Chapter 2) as well as an intrinsic control namely SNP fingerprint match. Because germline SNP fingerprint is the only way to correctly track serial samples from the same individual, we sorted the sequence data from all the serial samples using dbSNP database to identify the germline variants unique to each case. Within the 63 sequenced serial samples, 58 of them matched well with their corresponding samples and only 5 (Table 4.2) did not match this criteria, 3 of which had been received from outside the Bio-bank. Thus, 96.8% of the Bio-bank serial samples matched completely. The unmatched samples were excluded from the subsequent analysis.

Table 4.2. **An example of a number of germline SNPs used as fingerprints in tracking serial samples***

Variant information				CLL-1	CLL-1	CLL-1	CLL-9	CLL-9	CLL-16	CLL-16	CLL-16	CLL-19	CLL-19	CLL-19
Chr. Location	Ref.	Var	dbSNP	VAF%	VAF%	VAF%	VAF%	VAF%	VAF%	VAF%	VAF%	VAF%	VAF%	VAF%
2:141032088	C	T	rs1386356	100	44.1	49	100	100	100	100		100	100	100
2:141108531	G	A	rs16843826									100		
2:141259283	G	A	rs35296183	47.2				47.8					50.2	47.8
2:141260668	A	G	rs4444457	100	51.5	50	48.8	46.4		49.2			100	100
2:141274576	T	C	rs4954672	47.9	49	50	49.6	99	97.3	47.7	98.4	100	97.5	98.8
2:141274504	G	A	rs75124368	49.8										
2:141457985	T	A	rs13431727	42.8				47.9			43			
2:141274504	G	A	rs75124368	49.8										
2:141130695	C	T	rs16843864					47.7	46.4			100	46.6	47.7
2:141116420	C	T	rs150879175						47.9					
2:141116447	G	T	rs35546150							51.1				
2:141128779	C	G	rs76554185										52.1	48.1
2:141242918	T	C	rs34488772										53.3	52.3
2:141245204	T	C	rs74789055										51.8	50.4
2:198265526	A	G	rs788018	100	100	100	100	100	52.9	49.5	100	100	100	100
6:26157073	A	G	rs2298090								47.2			
7:82453708	A	C	rs2522833	50.8	98.2	100	46.7	47.5	43.6	46.1	95.3	97.8		
7:82544510	C	T	rs61995908									54.5		
7:82544987	A	G	rs17156844				50.7		98.6	51.9	53			
7:82581859	C	T	rs976714	51.1	51.2	57	44.2			49.1	99	49.3		
7:82582846	C	T	rs10954696	48.9	49.5	49	46.7			46.6	100	46.1		
7:82764425	C	G	rs2877	100	46.2	58	47.3	53.1	100	49.1	99	100		
7:82785097	T	C	rs61741659	45.9				46.2	45.9			46.3		
7:82435033	C	T	rs12668093		47.3	54			48.9					
SNP pattern match				x	✓		x		x	x		x	✓	

*Each variant (horizontal bars) labelled as wild type (white cells), heterozygous (green) and homozygous (yellow). Proposed serial sample(s) for each CLL case are labelled by (X) if no match or by (✓) if perfect match for SNP pattern was found with the last sample

4.3.2. AID-related mutations were associated with subclonal evolution

Defining mutational signatures is an important part of the analysis of the cancer genome to identify the operating mutational process [287]. Recently, distinct patterns of somatic mutations from WGS data of the initial CLL study (n = 4 by Puente et al) has revealed a high incidence of C>T changes in CLL [245]. A later study of unselected CLL cases (n = 28 by Alixandrov et al) decomposed the signature by analysing the nucleotide changes within the sequence context; thus provided a higher resolution of the derived mutational signatures [288]. In agreement with the former study, the later study revealed a predominance of C>T /G>A changes at CpG sites consistent with the physiological (ageing related) spontaneous conversion of cytidine to thiamine at CpG sites [289]. Moreover, a non-canonical cytidine de-aminase (nc-AID) enzyme-related signature characterised by genome-wide non-clustered A>C/T>G changes at WA/TW sites (where W = T or A) was identified. This signature is produced by downstream error prone polymerase ɳ enzyme function. In addition, APOBEC3 signature characterised by C>T/G>A at TCW/WGA motifs was found [290]. A subsequent study of 30 CLL samples at a single time point by Kasar et al (published in 2015) revealed canonical (c-AID) mutational signature. This signature is characterised by C>T /G>A at WRCY /RGYW (where W = A or T, R = purine, Y = pyrimidine) motifs [137]. The production of such signature is associated with restorative function of downstream Uracil DNA glycosylase enzymes (UNG) which usually occur as regionally clustered mutations. Furthermore, based on clonality of the mutations determined by VAF%, it has been proposed that c-AID-induced mutations possibly occurred before and after the latest selective sweeps. As both ageing (which is an ongoing process) and c-AID-related mutations had clonal and subclonal fractions, therefore a possibility of ongoing c-AID activity was suggested [290].

Here, our NGS data from 23 CLL samples at different stages of the disease provided an opportunity to test this hypothesis. Initially, the acquired somatic non-synonymous substitutions in the 23 tested samples taken at latest time points were examined for mutation signatures. Of these samples, only 21 carried single nucleotide mutations (the other 2 samples carried only indels). The 60 single nucleotide mutations in the 21 samples were examined using their immediate nucleotide sequence context. For each point mutation, we analysed three nucleotides, including the two at immediate 5’ and 3’ sides of the mutated nucleotide. Thus each of the 6 possible mutations had 16 triplets, giving a total

number of 96 somatic substitutions in Watson and Cricks orientation [290]. In our analysis of the results for samples taken at the latest time point, C:G > T:A substitutions (n=18) were found to be highest, accounted for 30.0% of all point mutations (Figure 4.1.A). We then examined the nucleotide substitution profile between the samples before and after treatment in 14 cases. It was found that C>T changes increased from 28% to 36% (Figure 4.1.B and C).

Although C>T changes can be induced by AID, they may also associate with ageing and other factors, e.g. APOBEC3 family members. We therefore examined in detail their mutational signatures in the latest samples of the 21 cases described above. 9 of the 60 mutations were identified as ageing signature (C >T mutations at CpG sites), 8/60 were AID-related signature and 44/60 were others. Next, we questioned whether there is any difference between the evolution of ageing-related, AID-related and other factors mutations throughout the course of CLL. We compared other aspects, including age of the patients and time interval between sampling. There was no significant difference in age (P = 0.882) or the time interval (P = 0.915) among these three signatures (Table 4.3). However, as shown in Table 4.4, we found that, 7 out of the 8 (85.7%) of AID-related clonal mutations showed the evolution (that is increase in the VAF) while the remaining one showed clonal stability.

Similarly, 26/43 (60.47%) of other factor-related mutation clones showed evolution. While only 2 out of the 9 (22.22%) ageing-related mutation clones had an increase in the VAF. This showed a clear difference among the three groups (P = 0.021, χ^2 test). In addition, the clonal expansion in both AID and other factor-related mutations reached statistical significance (p = 0.018 and <0.001, respectively, Wilcoxon test). In contrast, there was no statistical difference in clonal size in the follow-up in patients with ageing-related mutations (p = 0.484, Wilcoxon test).

Table 4.3. **Ages and follow-up in patients with different mutation signatures**

Variables	Mutation signatures		
	Ageing	AID	Others
Ages (median (range))	66 (55-82)	65.5 (37-76)	70.5 (55-79)
	p = 0.882 (ANOVA test)		
Time intervals (median (range))	32 (7-61)	26 (18-60)	44 (4-61)
	p = 0.915 (ANOVA test)		

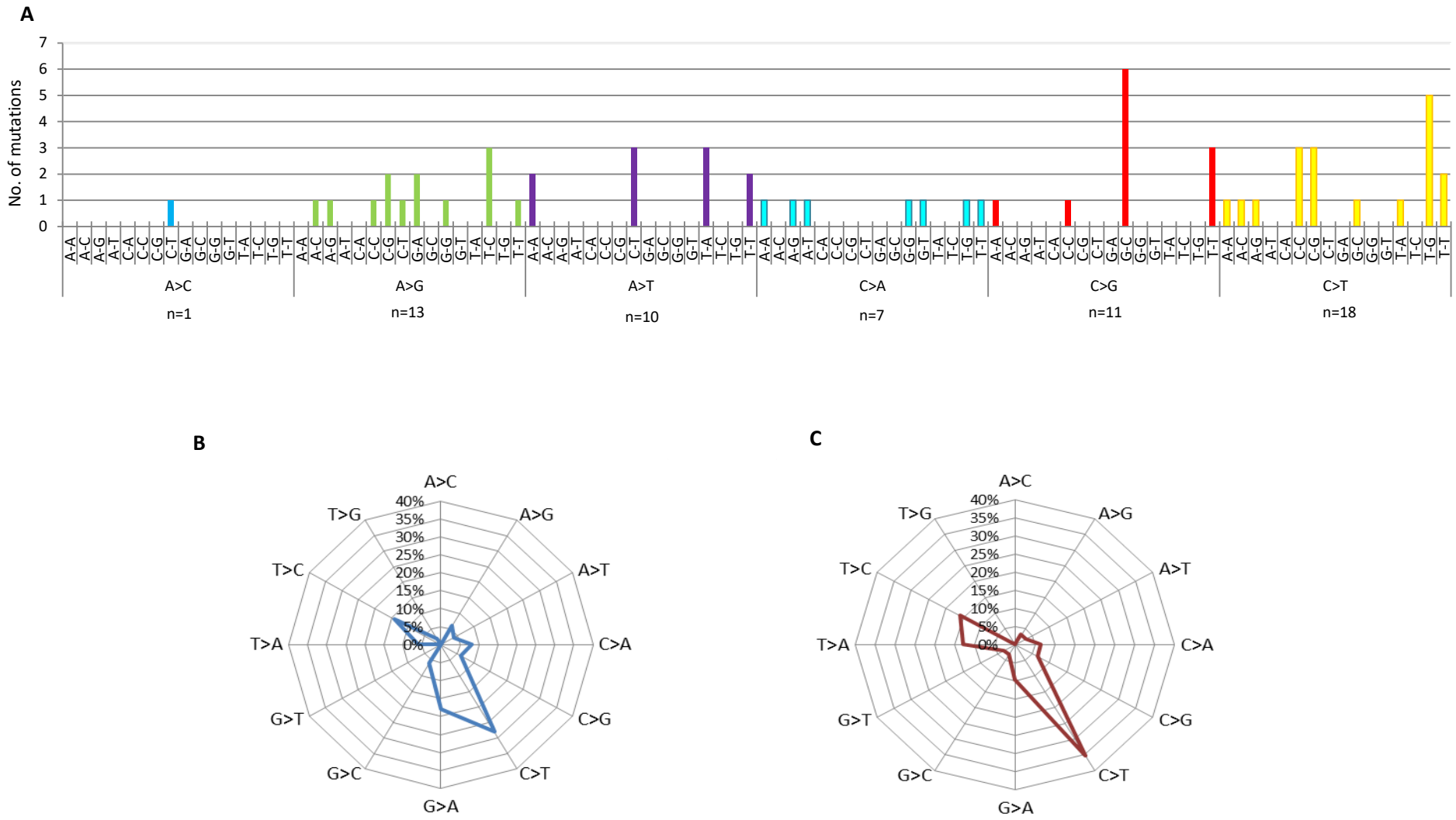


Figure 4.1. **Analysis of somatic non-synonymous single nucleotide substitutions**

Distribution of nucleotide substitutions in tri-nucleotide sequence context in 21 CLL samples with point mutations (**A**). % of various nucleotide substitutions in 14 CLL cases before (**B**) and after therapy (**C**).

Table 4.4. **Summary of different signatures VAF% in earliest and latest CLL samples**

Signatures	Case	Gene	Earlies sample VAF%	Latest sample VAF%	Time interval (months)
Aging related	CLL-2	*XPO1*	29	25	18
	CLL-3	*SF3B1*	5	4	60
	CLL-5	*TP53*	65	84	7
	CLL-8	*SF3B1*	4	2	40
		PCLO	15	4	
	CLL-12	*LRP1B*	3	0	61
	CLL-13	*XPO1*	36	24	55
	CLL-20	*SF3B1*	0	5	11
	CLL-23	*SAMHD1*	49	49	24
AID related	CLL-1	*LRP1B*	0	6	18
	CLL-2	*TP53*	15	40	18
		PCLO	49	49	
	CLL-3	*TP53*	1	43	60
	CLL-10	*NOTCH1*	3	21	24
	CLL-15	*ATM*	14	18	28
	CLL-21	*SF3B1*	22	41	44
	CLL-26	*HIST1H1E*	39	48	74
Others	CLL-1	*SF3B1*	1	6	18
			4	14	
		PCLO	0	6	
		TP53	3	10	
			2	4	
			7	25	
			1	3	
	CLL-2	*LRP1B*	3	0	18
	CLL-3	*ATM*	0	6	60
	CLL-4	*SF3B1*	49	49	4
		TP53	74	74	
	CLL-5	*SF3B1*	40	47	7
		PCLO	36	36	
	CLL-6	*SF3B1*	45	53	52
		TP53	4	1	
			4	96	
	CLL-7	*FBXW7*	0	10	48
		TP53	36	31	
			14	54	
	CLL-8	*TP53*	2	11	40
			1	0	
	CLL-10	*ATM*	54	51	24
		LRP1B	0	4	
		SF3B1	2	5	
	CLL-11	*ATM*	37	40	40
			38	37	
	CLL-12	*SF3B1*	10	30	61
		ATM	98	98	
	CLL-13	*ATM*	11	10	55
	CLL-14	*PCLO*	39	37	34
		ATM	0	6	
		SAMHD1	58	72	
	CLL-15	*ATM*	0	6	28
	CLL-17	*ATM*	41	48	61
		SF3B1	8	2	
	CLL-19	*FBXW7*	5	4	10
		CHD2	21	28	
		CHD2	21	30	
	CLL-20	*SF3B1*	55	80	11
	CLL-22	*LRP1B*	48	49	34
		ATM	4	0	
	CLL-23	*MYD88*	4	0	24
	CLL-26	*PCLO*	44	59	74

4.3.3. Ageing-related mutations were associated with UM-*IGHV* in CLL

Following the above observations, we questioned whether there is any difference in these signatures between *IGHV* mutated and un-mutated cases. When calculating and comparing the frequency of the occurrence of each type of these signatures in each *IGHV* group, we found that C:G>T:A and C:G>G:C substitutions are more frequent in *IGHV* unmutated cases (Figure 4.2.A). Furthermore, all ageing-related mutations were presented in this group (Figure 4.2.B). This frequency (46.15%, 6/13) was obviously different from that (0) in cases with M-*IGHV* (P = 0.051). However, there was no statistically significant difference in AID- or other factor-related mutations between *IGHV* mutated and unmutated groups (P = 0.612 and 0.356, respectively, Fisher's exact test).

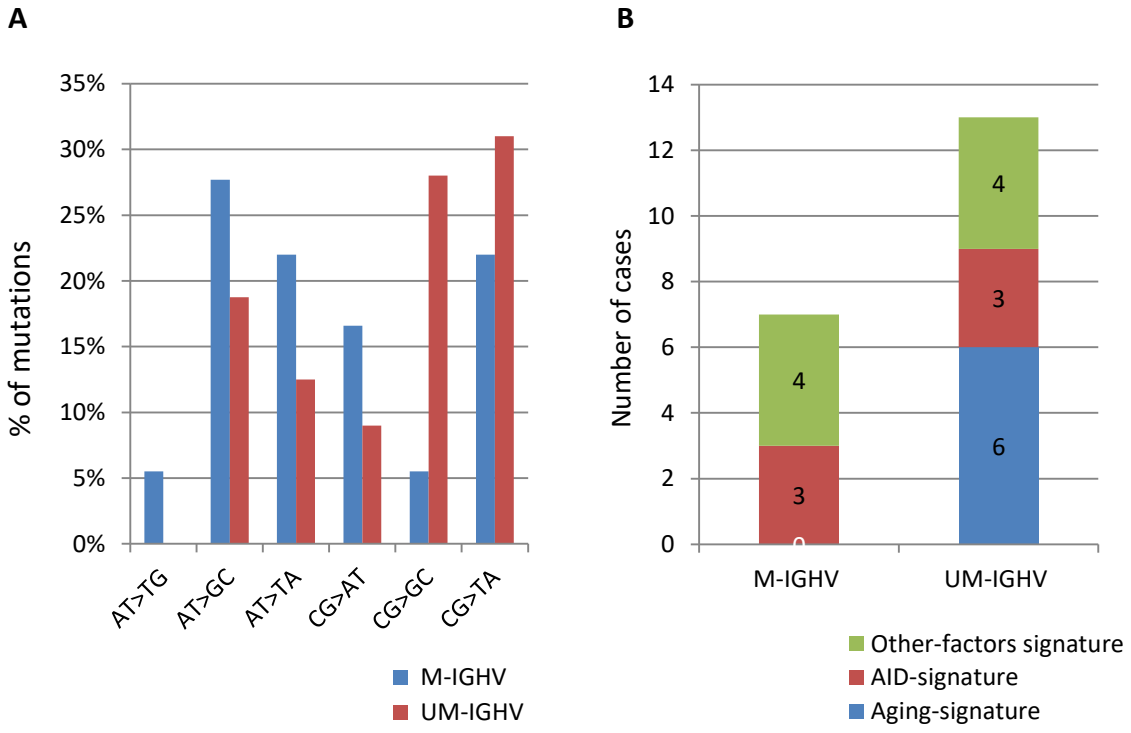


Figure 4.2. **Comparison of single nucleotide substitutions and mutation signatures between patients with M- and UM-*IGHV***

A. Frequency of each type of single nucleotide substitution is shown in this histogram plot. **B.** Distribution of number of case with different mutation signatures in the two groups.

4.3.4. Evidence of complex subclonal architecture before and after disease progression and its persistence in relapse

Recently, a number of CLL studies applied high throughput, but less sensitive techniques (e.g. WGS and WES) to define molecular heterogeneity and clonal/ subclonal architecture at various time points. Hence, small mutated subclones might not be detected. As a result, the definitive relevance of subclonal heterogeneity in CLL development has not been fully understood yet. Therefore, it was obviously worth to track the subclones with recurrent mutations throughout CLL progression and/or relapse by ultra-deep and highly sensitive sequencing, as planned in this chapter.

As shown in Figure 4.3, 89.3% of the mutated clones/subclones identified at the latest time point were detectable at time of diagnosis or prior to treatment. Specifically, *TP53* mutated subclones existed prior to treatment and became dominant after treatment (CLL-1, CLL-2, CLL-3 and CLL-6). In contrast, subclones with *ATM* mutations or deletions were mostly stable at both early and late stages. However, the *ATM* mutations in CLL-14 and CLL-15 clearly occurred later than the background genetic event (11q-, as detected by FISH). In 9/11 of the *SF3B1* mutated samples, co-existing *ATM* or *TP53* mutations were found. In the evolutionary process, almost invariably co-existing subclones carrying *SF3B1* mutations were overcome by clones carrying *ATM* or *TP53* mutations. While sole *SF3B1* mutations were present and evolved over disease progression and at relapse in CLL-20 and CLL-21, highlighting the possible role of this gene in CLL progression and relapse.

Like the subclones in the above mutated genes, all the remaining subclones with mutations in other genes showed expansion over time of follow-up until the VAF reached close to 100% or around 50%, depending on whether or not the wild-type allele was deleted (Sections 3.3.11 and 4.3.8). However, the number and type of mutations and genes with such mutations in these subclones remained unchanged and preserved throughout the course of the disease.

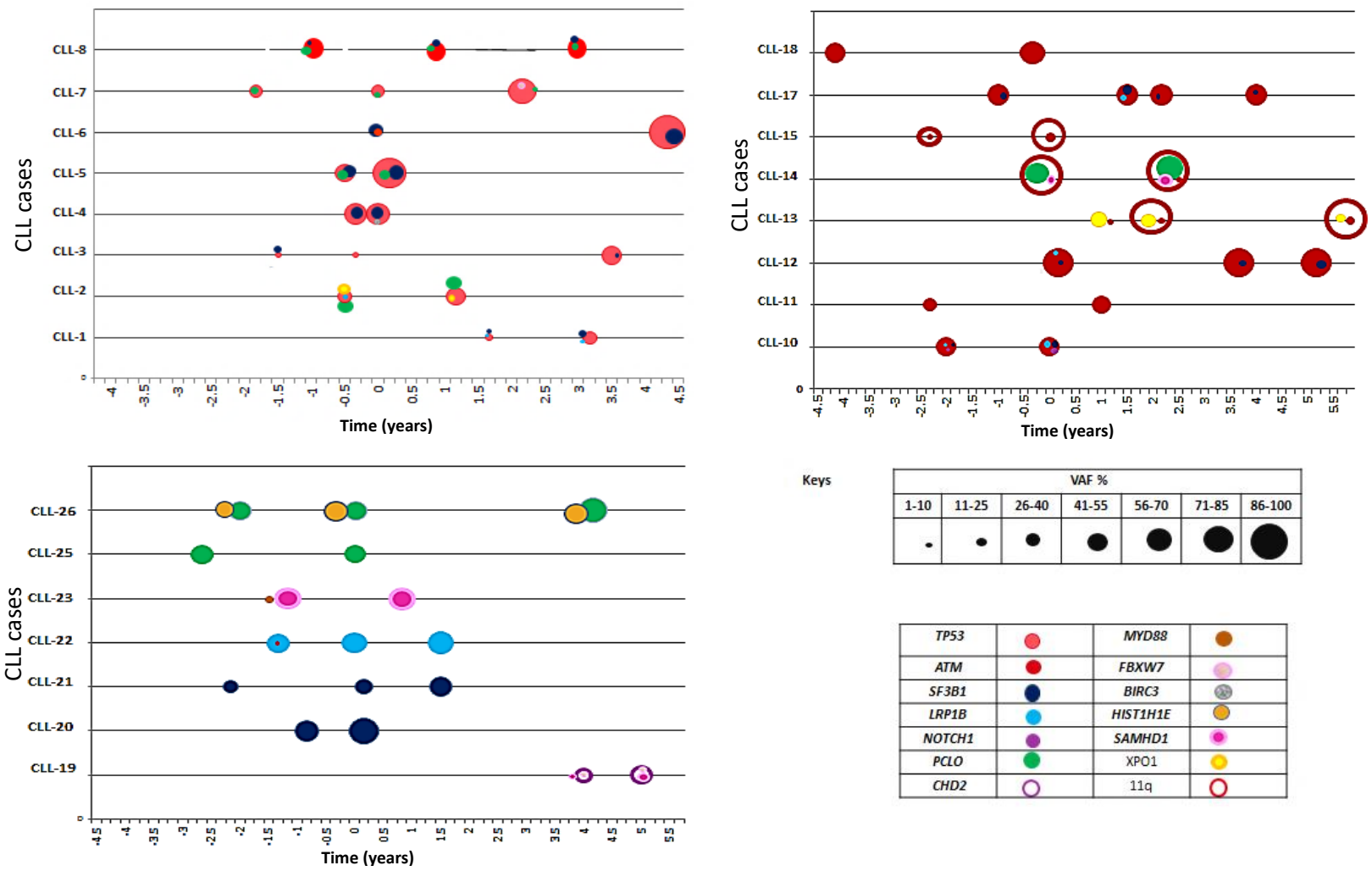


Figure 4.3. **Mutation dynamics and subclonal architecture of 23 CLL cases with progressive disease**
Serial samples taken before (-) and after (+) the first treatment (0) were screened for the recurrent mutations. For genes with multiple mutations, the mutation with highest VAF% is presented to show clonal size

4.3.5. Mutations in *TP53*, *SF3B1*, *NOTCH1* and *BIRC3* demonstrated faster clonal evolution

To assess the contribution of each targeted gene mutation to disease progression, we compared the VAF doubling time of these mutations. This will provide useful information of risk of each gene alteration for CLL disease progression.

We reasoned that highly fit CLL subclones grow faster over time of evolution; mutations related to high fitness can be characterised by rapidly increasing mutant allele frequency within the tumour. Though the fitness potential is not measurable and that the rate of evolution is not the sole indicator of fitness, it is at least a sign of strength. Given this, the speed of mutant allele frequency doubling for each gene was calculated. We noted that the magnitude of expansion is not the same for subclonal and clonal mutations; mutations found to be clonal at the earliest time point did not show further expansion at time of follow-up. This is because the maximum limit of extension was already reached; therefore we excluded all the clonal genetic lesions in this analysis.

As expected, there were clear differences in the doubling time among different mutated genes. A shorter median doubling time of < 2.5 years was observed for mutations in *TP53*, *SF3B1*, *NOTCH1* and *BIRC3* as compared to other targeted genes (the medians ranged from 0.16 to 2.2 year) (Figure 4.4.A). Therefore, mutations in these genes were likely to drive higher fitness or more aggressive clones which might result in progression of CLL. According to the most recent risk classification of CLL, these four genes can be classified as high risk, while *ATM* as intermediate risk, for which we also found a longer median doubling time (10.95 years). We grouped the gene mutations according to the above proposed risk classification [291] into high risk, unknown risk and intermediate risk. Statistical analysis revealed significant differences between the 3 groups (P = 0.01, ANOVA test). Further analysis confirmed the significance of difference between the doubling time of the high risk and the intermediate risk genes (p = 0.008, Mann-Whitney test) while no difference for unknown risk gene mutations (*CHD2*, *LRP1B*, *SAMHD1*, *HISTH1E* and *PCLO*) (median doubling time 2.55 year) as shown in Figure 4.4.B. Therefore, for risk stratification of this group of genes, it might be possible to be in between the high and intermediate group.

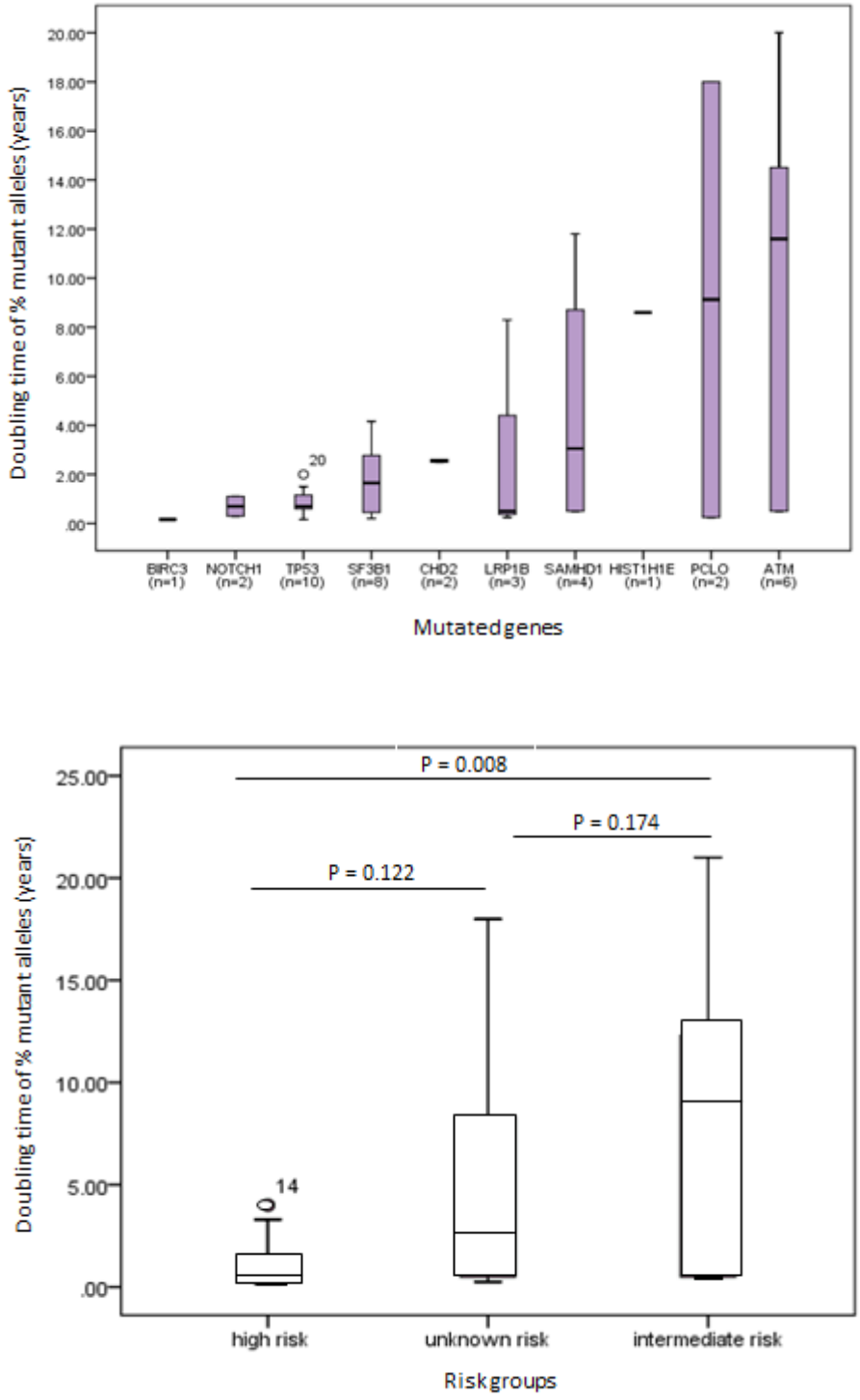


Figure 4.4. **Doubling time of mutant alleles in the targeted genes**
The doubling time was calculated based on VAF (%) of each mutation detected at earliest and latest time points for the 23 patients as shown in Figure 4.5, and presented as median and quadrants for **A**. each mutated gene. **B**. for grouped mutations with known and unknown risk categories (high risk: *TP53*, *SF3B1*, *NOTCH1* and *BIRC3*; intermediate risk: *ATM*; and unknown risk: *CHD2*, *LRP1B*, *SAMHD1*, *HIST1H1E* and *PCLO*).

4.3.6. Evidence of growth priority for subclones with deleterious mutations

Generally, the process of positive selection of somatic mutations contributes to tumour plasticity in cancer. In our longitudinal study, 20 out of the 23 cases harboured more than one mutated subclones. When focusing on the dominant mutations, we found that the most commonly affected genes in the early collected samples were *TP53* and *ATM*, each of which was detected in 5 cases. However, the number of mutated *TP53* increased to 8 in samples taken at the late stages. In addition, *NOTCH1* overtook *ATM* as the dominant mutated gene in CLL-10. The numbers of other dominant mutated genes were either not changed (in *ATM*, *SAMHD1*, *PCLO*, *LRP1B* and *CHD2*) or reduced (in *SF3B1* and *XPO1*) (Table 4.5 and Appendix 7.3.5). More interestingly, we also found convergent subclonal evolution in CLL-7 and CLL-11. Multiple mutations were identified in *TP53* and *ATM* in the former and the latter, respectively, and one of them became the dominant mutation.

To understand the possible roles of dysfunction of those affected genes in the subclonal evolution, we defined deleterious mutations in the order: truncating or shift of reading frame > missense mutations; and in case of missense mutations in *TP53*, non-functional > partially functional as estimated based on its transcriptional activities (TA) in the database of International Agency for Research on Cancer (IARC) TP53 (http://www-p53.iarc.fr/) [292]. In the early sample taken before chemotherapy from CLL-11, there were two subclones (A and B) both with a separate mutation in *ATM*, with 37% and 38% VAF, respectively. The mutation in subclone A was deleterious and resulted in truncation of the gene at exon 10, while the one in another subclone was a missense mutation close to the C terminal (exon 57). Not surprisingly, subclone A with the deleterious mutation expanded to dominate over subclone B (VAF 40% versus 37%) after the treatment (Figure 4.5.B).

Likewise, CLL-7 had two different *TP53* mutations within the DNA binding domain with changeable VAF in the three serial samples tested. As presented in Figure 4.5.D, the missense mutation (p.C135W) that partially retained TA of p53 (median (range): 38.45 (21.0 - 79.6)) was initially the dominant mutation and showed an expansion together with the increase of circulating lymphocyte count and lymph node size (Figure 4.5.C) before chemotherapy. However, it declined after the treatment. In contrast, another subclone with a different missense mutation (p.R213L),resulting in total loss of the TA, (median(range):

00.00 (0 - 3.8)) markedly expanded with the VAF increasing from below 15% before treatment to over 54% in less than 2 years of follow-up after the treatment.

Table 4.5. **Comparison of dominant mutated genes and clones between the earliest and latest collected samples in cases with ≥ 2 mutations**

Cases	Dominant gene (clone)		Convergent clonal evolution
	Earliest	Latest	
CLL-1	*TP53* (C)	*TP53* (C)	
CLL-2	*PCLO* + *XPO1*	*PCLO* + *TP53*	
CLL-3	*SF3B1*	*TP53*	
CLL-4	*TP53*	*TP53*	
CLL-5	*TP53*	*TP53*	
CLL-6	*SF3B1*	*TP53*	
CLL-7	*TP53* (A)	*TP53* (B)	Yes
CLL-8	*TP53* (A + B)	*TP53* (A + B)	
CLL-10	*ATM* (A)	*ATM* (A)	
CLL-11	*ATM* (B)	*ATM* (A)	Yes
CLL-12	*ATM*	*ATM*	
CLL-13	*XPO1*	*ATM*	
CLL-14	*SAMHD1* (B)	*SAMHD1* (B)	
CLL-15	*ATM* (A)	*ATM* (A)	
CLL-17	*ATM*	*ATM*	
CLL-19	*CHD2*	*CHD2*	
CLL-20	*SF3B1* (A)	*SF3B1* (A)	
CLL-22	*LRP1B*	*LRP1B*	
CLL-23	*SAMHD1*	*SAMHD1*	
CLL-26	*PCLO*	*PCLO*	

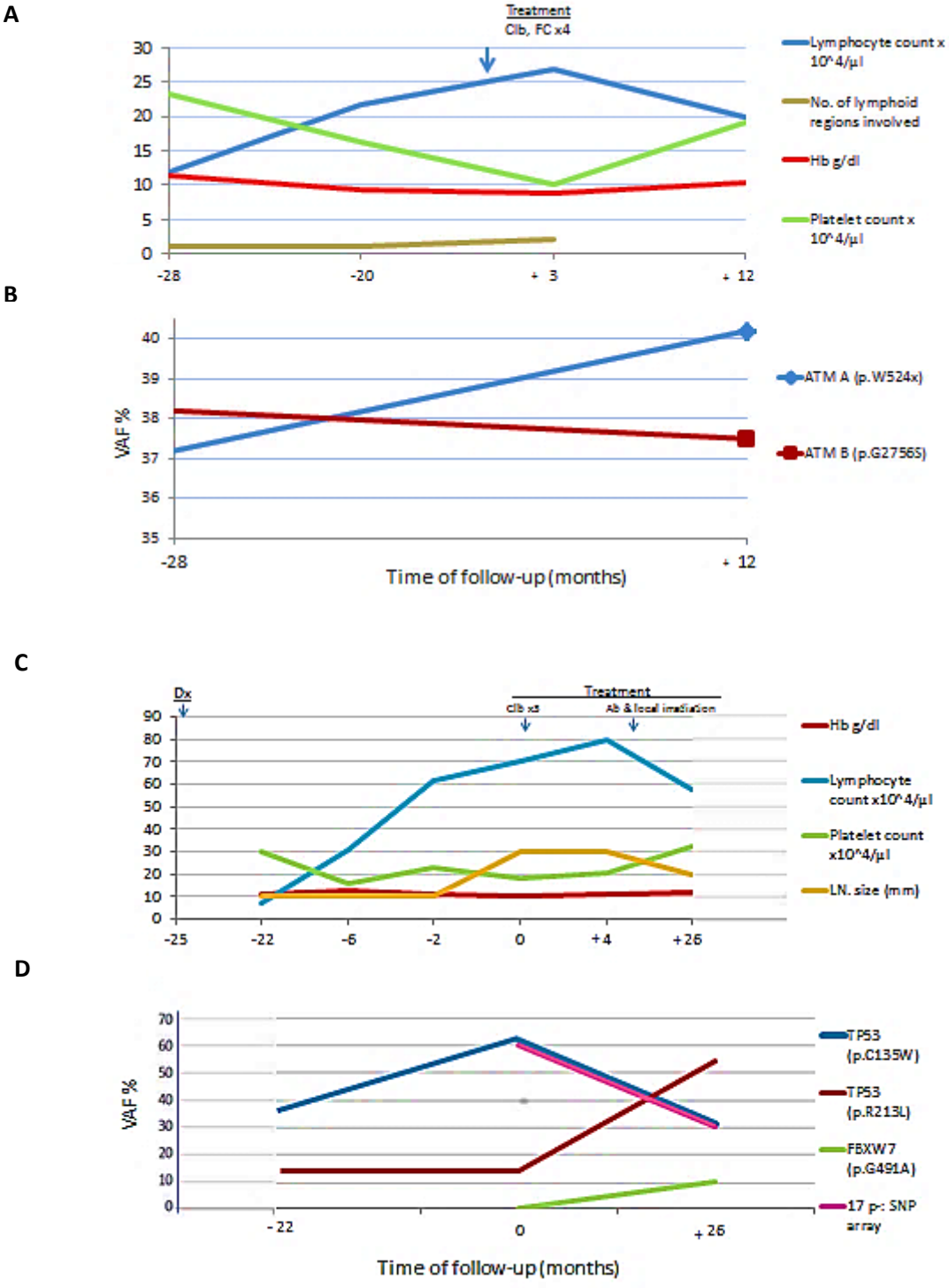


Figure 4.5. **Convergent clonal evolution favours deleterious mutations**
Clinical information available for longitudinal samples of CLL-11 (**A**) and CLL-7 (**C**) at various time points shown in months before (-) and after (+) the first treatment (0) is shown. NGS results show evidence of growth priority for deleterious mutations of *ATM* and *TP53* in a background mixture of multiple independent point mutations in CLL-11 (**B**) and CLL-7 (**C**), respectively.

4.3.7. Linear clonal evolution prevailed the patterns of evolution

By definition, the branched evolutionary path in clonal evolution is a replacement of parent clones by the progeny in a single or multiple successive full selective sweeps, while the linear evolutionary trajectory is co-existence of multiple subclonal populations which compete for growth [286]. Therefore, one can predict the shape of evolution from tumour subclonal components before and after disease progression and/or relapse from the variant allele frequency and the trend of change in sequential samples [201].

To address the question of which pattern of evolution is predominant in our cohort, we analysed the subclonal mutation profiles in these serial samples taken at the earliest and latest time points. In only 6 of the 23 cases we found gain of additional subclonal mutations (Table 4.6). Moreover, none of these newly emerged subclones became dominant in the follow-up. While 5 small mutations (VAF 1% to 4%) in *TP53*, *ATM*, *LRP1B* and *MYD88* in CLL-8, CLL-22, CLL2, CLL-12 and CLL-23, respectively, became undetectable. In contrast, 20 of the 23 cases showed clonal expansion from the same subclones previously identified as shown in Figure 4.4. These observations clearly showed that the linear clonal evolution pattern was dominant within our cohort.

Table 4.6. **Timing of additional subclonal gene mutations gained in 6 CLL cases**

	CLL-1	CLL-7	CLL-10	CLL-14	CLL-15	CLL-20
Genes	*LRP1B* and *PCLO*	*FBXW7*	*LRP1B*	*ATM*	*ATM*	*SF3B1*
Time	After relapse	After treatment	After disease progression	After treatment	After treatment	After treatment

4.3.8. Longitudinal copy number analysis revealed clonal evolution in a CLL patient

As mentioned in Section 4.2.1, 3 serial samples belonging to CLL7 taken at time of diagnosis, at time of disease progression (before treatment) and after treatment (when the disease relapsed) were subjected to CytoSNP 850K array. The retrieved raw data was checked for quality control. As data from the sample at diagnosis did not pass the quality control, therefore it was excluded from subsequent analysis.

We observed complex karyotype before and after treatment with chromothripsis-like event affecting chromosome 6 in both analysable samples. Compared to the sample before treatment, the post treatment sample acquired CNA involving 5 additional bands of 3 chromosomes (Chromosomes 10, 11 and 19) (Table 4.7).

Table 4.7. **Evolution of chromosomal copy number changes in a CLL case studied** •

Case	Before treatment					After treatment				
	Start Cyto	End Cyto	Start	End	Size (bp)	Start Cyto	End Cyto	Start	End	Size (bp)
CLL-7	2p25.3	2p21	14,238	46,082,995	46,068,757	2p25.3	2p16.1	14,238	58,795,436	58,781,199
	2p16.3	2p16.1	48,771,605	59,409,979	10,638,375	2p13.2	2p11.2	73,361,948	83,633,730	10,271,783
	2p13.1	2p11.2	73,556,155	84,249,446	10,693,292	6p25.2	6p24.3	3,077,141	9,275,594	6,198,454
	6p25.2	6p24.3	3,205,325	9,745,037	6,539,713	6p22.3	6p22.3	16,998,804	19,017,212	2,018,409
	6p21.33	6p21.2	30,724,430	38,723,575	7,999,146	6p21.33	6p21.2	30,770,000	38,728,142	7,958,143
	6p21.1	6q12	44,568,740	66,916,257	22,347,518	6p21.1	6p11.2	44,509,266	58,630,693	14,121,427
	6q13	6q15	74,217,278	88,021,118	13,803,841	6q11.1	6q12	61,891,118	67,977,326	6,086,209
	7q21.11	7q21.12	80,512,536	86,608,786	6,096,251	6q13	6q15	74,166,098	88,152,535	13,986,438
	7q31.1	7q32.1	112,044,527	127,479,262	15,434,736	7q21.11	7q21.12	79,585,451	87,226,690	7,641,240
	7q32.3	7q36.3	131,304,940	159,126,310	27,821,370	7q31.1	7q36.3	112,071,795	159,126,310	47,054,516
	8p23.3	8p11.23	164,984	36,646,217	36,481,234	8p23.3	8p12	164,984	35,042,190	34,877,207
						10q24.1	10q26.3	99,267,639	135,477,883	36,210,245
						11q22.3	11q25	107,617,257	134,934,063	27,316,807
	17p13.3	17p11.2	12,344	21,246,375	21,234,032	17p13.3	17p11.2	12,344	20,117,151	20,104,807
						19p13.3	19p13.11	5,830,302	18,342,477	12,512,176
						19p13.11	19p11	18,347,539	24,487,350	6,139,812

• The blue filled CNAs denote gains and the white filled CNAs are losses

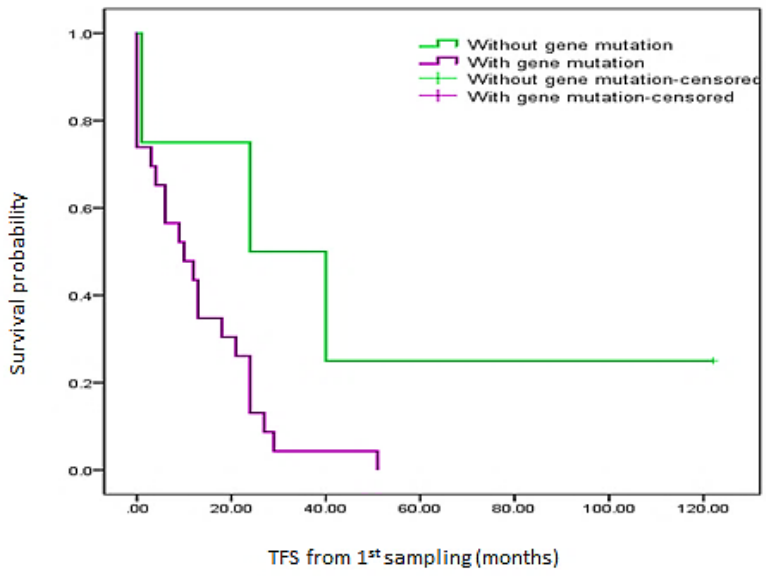
4.3.9. Clinical impact of mutations detected at early stages of CLL

To understand whether these mutations have predictive value for adverse clinical outcome in CLL, we studied the relationship of mutations detected in samples taken at the earliest stage (10 to 65 months before of the latest sampling) (Figure 4.3). When compared with the

4 having no mutations (CLL 29, 30 31 and 32, in Table 3.6.), the 23 patients with mutations in the earliest samples had a similar OS (measured either from time of diagnosis or the time of 1st sampling). This was the same in a comparison among cases with no mutation, non-driver mutation and driver mutation (data not shown). Similar to what found in Chapter 3, there was no difference in the TFS measured from the diagnosis between patients with and without the mutations (data not shown). However, when it was measured from time of the 1st sampling, the TFS became clearly shorter in patients with the mutations (Median TFS: 10 months versus 24 months), although the P value (0.061) was above the pre-set α level (0.05) (Figure 4.6.A). Moreover, by dividing the former group into those with and without driver mutations, we could not find a difference in TFS measured either from the diagnosis or time of the 1st sampling between them. This prompted us to explore whether the size of subclones with driver mutations had an impact on the TFS.

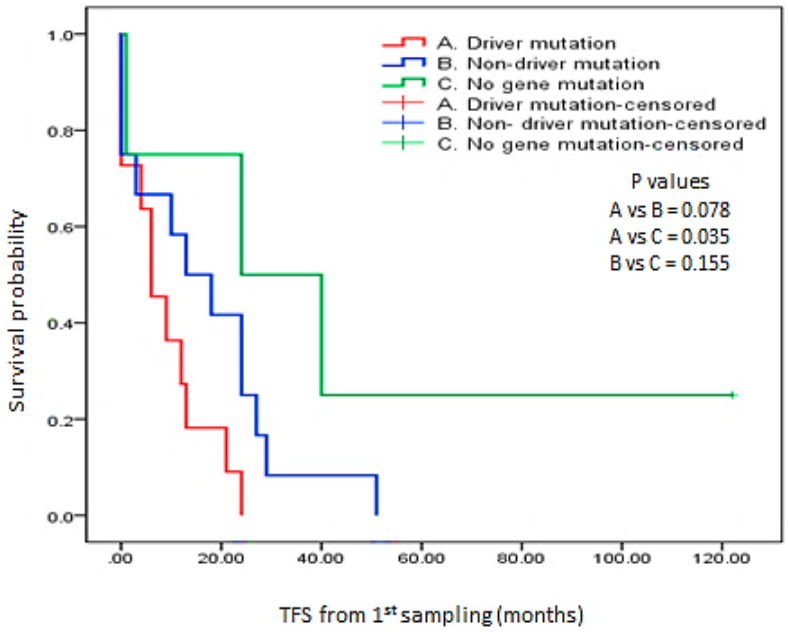
To do so, we used the criteria reported by Davide Rossi et al [293] to select the significant size of all of the five driver mutated genes detected in our cohort: *BIRC3* $\geq$ 1%, *SF3B1* $\geq$ 35%, *NOTCH1* $\geq$ 25%, *SAMHD1* $\geq$ 7% and *TP53* $\geq$ 1%. As expected, a statistically significant difference in the TFS measured from time of the 1st sampling was observed among the three groups of CLL patients (P = 0.037, Log-rank test), with those harbouring a large driver mutation showing the shortest TFS (Figure 4.6.B).

A



Groups of CLL patients	No. of samples	Median TFS	95% CI	P
Without gene mutation	4	24.0	0.00 - 62.22	0.061
With gene mutation	23	10.0	0.61 - 19.39	

B



Groups (code)	No. of samples	Median TFS	95% CI	P
Driver mutation (A)	11	6.0	0.60 - 11.39	0.037
Non-driver mutation (B)	12	13.0	0.00 - 26.57	
No gene mutation (C)	4	24.0	0.00 - 62.22	

Figure 4.6. **Effects of target gene mutations detected at an early stage on TFS measured from the time of sampling**

A. Kaplan-Meier curve and Log-rank test showing a trend in which patients with the mutations had a shorter TFS than those without them. **B.** Kaplan-Meier curve and Log-rank test showing statistically significant differences in the TFS between cases with and without the large-size driver mutations.

4.4. Discussion and conclusions

In Chapter 3, we have clearly shown that the majority of CLL patients carry at least one of the recurrently mutated genes when tested at the stage of disease progression and/or chemotherapy resistance. In this chapter we aimed to investigate the evolution of these mutant subclones in the course of the disease. Other studies have shown evidence of complex subclonal architecture and its association with disease progression and relapse in CLL. However, with a limited coverage depth and therefore a low sensitivity of either whole genome or whole exome sequencing approaches [198], they might have underestimated small subclones and thus had not provided a complete picture of subclonal composition. However, NGS approaches targeting only the coding sequence of a limited number of genes such as *TP53* has massively increased the sensitivity [273]. They provide a useful tool to study the more comprehensive subclonal architecture and its dynamics. This would possibly re-shape the portrait of intra-tumour subclonal complexity at various stages of the disease and depict the effects of treatment on the evolutionary dynamics. We therefore applied such an approach to investigate the clonal evolution of those recurrent somatic non-synonymous mutations and its relations with clinical outcomes of patients with CLL. Here, we described the results of the longitudinal study using sequential samples taken at different time points from patients with mutations detected in Chapter 3. Use of the same deep sequencing approach allowed us not only to track minor subclones, but also to collect data that were suitable for comparison with those produced using samples taken at the late disease stage and sequenced in Chapter 3. For example, the SNP patterns matched well in serial samples from each individual case. This was successfully used as a genomic fingerprint to reconfirm the real source of those serial samples and guaranteed the truthfulness of results from the paired samples presented in this chapter.

After this, we investigated the mutation signature both archetypically and contextually within tri-nucleotide sequences. This topographic information was important for defining the operating mutational process before and after disease progression and/or relapse in CLL in general, as well as in different subgroups with mutated or unmutated *IGHV*. Using mutation signature information based on the type of nucleotide changes and sequence context, we were able to identify three main mutational signatures to be operative in CLL, including AID- and ageing-, and other factor-related signatures.

AID is the key enzyme contributing to the immunoglobulin somatic hypermutation process in B-cell development and is responsible for affinity maturation of antibodies [294]. Its enzymatic activity is tightly regulated to target IGV region genes by catalysing cytosine deamination to produce uracil, thus creating U: G mismatches [294]. It has been suggested that off-target (aberrant) activity of AID may contribute to oncogenic mutations outside IGV and chromosomal instability [290]. In this respect, the AID mutational signature has been very recently identified in an unsupervised analysis of mutational signatures for 30 unselected CLL cases at a single time point using WGS [290]. Here, in analysis of mutation signatures recorded in the longitudinal study, we are the first to find that AID-related mutations were most likely associated with subclonal evolution in CLL. Our finding supports the hypothesis proposed in a recent study suggesting the possible role of AID activity in evolution of mutated subclones in this disease [290]. Notably, AID seems to be not the only factor that contributes to this, as clonal expansion was also observed in our study for mutations not related to this enzyme. These non-AID-related mutations may relate to ageing and other factors. To support this, 2 of the 9 ageing-related mutations identified by us had subclonal expansion in the disease course, although it was not as frequently as AID-related mutations. In addition, the ageing-related mutations occurred more frequently in *IGHV*-unmutated patients, suggesting its association to tumour cells with higher replicative capacity as reported by others [295] and therefore supporting the notion that unmutated *IGHV* have greater proliferative capacity and worse prognosis [296]. Taken together, these findings decipher the independent mutational signatures operative in these samples. The ongoing activity of AID might have an influence on clonal evolution in the off-target genes.

Our study did not provide direct evidence for specifying any other factors. However, other members of the APOBEC3 family, in which AID is one, may involve in the subclonal evolution. This is because they induce the same initial nucleotide changes as AID does [297].

Considering the mutation spectrum and subclonal complexity, we clearly showed a complex subclonal architecture of *TP53* and *ATM* mutated patients from early stages of the disease (Figure 4.3). An important question addressed here is whether difference in clonal evolution pattern is due to different pathways which are affected by the mutations. Here we reported that *TP53*, *SF3B1, BIRC3* and *NOTCH1* mutations were associated with the fast expansion of the subclones. It is well known that p53 as a gatekeeper at cell cycle check points is vital in

regulation of DNA repair, in addition of other functions, e.g. activation of apoptosis. In CLL, it has been found that genomic aberrations of *TP53* and its upstream regulator *ATM* are frequently enriched in patients with resistance to DNA damaging agents [137]. Moreover, these genetic lesions allow acquisition of additional genomic alterations contributing in CLL phenotypic transformations [191].

Considering other driver mutations, *SF3B1* (a recurrently mutated gene in CLL) has been found to be associated with clonal evolution [298]. It has also been reported by Schwaederlé et al in 2013 using Sanger sequencing technique that *SF3B1* mutation expansion could occur with or without therapy administration [299].

Moreover, subclonal mutations in other drivers such as *NOTCH1*, *BIRC3* have been linked to faster disease progression and poor survival [300]. We observed a similar trend related to *NOTCH1 and BIRC3* mutations, both of which were associated with a fast subclonal expansion (Figure 4.4). Despite of the requirement for a bigger number of samples for a solid conclusion, this finding provided a further support for the notion that subclonal mutations in *BIRC3* and *NOTCH1* can predict poor prognosis as these lesions could quickly evolve over time [138, 273].

Regarding the cause of growth priority in convergent clonal evolution (defined by presence of independent somatic mutations in the same gene) that has been observed in previous longitudinal studies of CLL in which *SF3B1*, *NOTCH1*, *ATM*, *TP53* and *BIRC3* genes were affected [201, 273]. Here, we documented the convergent clonal evolution in two cases with mutated *ATM* or *TP53.* We were able to use mutation types and their causal impact on wild-type protein functions (for p53 only) to explore their relationship with the subclonal evolution. We found that the truncating mutation at the N-terminal in *ATM*, and the non-functional mutation in *TP53* became predominant over the missense at the C-terminal and partially functional mutations in the same two genes, respectively, after chemotherapy. This finding emphasised an importance of these deleterious mutations in the subclonal selection process, which may involve competition over growth conditions in the micro-environment and adaption to extrinsic pressure such as chemotherapy [301].

To infer temporal order of gene mutations and the pattern of clonal evolution, phylogenetic trees are often employed by assuming that common predecessors are early events, but this method typically needs larger number of alterations than the ones available in our cohort [302]. Furthermore, standard statistical techniques have extremely limited power to assess the significance of different branches using longitudinal data from patients in few selected driver genes. Therefore, we used similar approach used by P. Quillette et al, published in 2013 [286] and depicted the pattern of evolution based on the presence or absence of the mutations and the trend of evolution during the time of follow up. In agreement with the models proposed [242], our study identified both linear and branch clonal evolutions in CLL, however, the majority of these mutant subclones followed a linear evolution, while branching evolution was rare in our cohort. The reasons for this are not yet known, but it might be due to deeper and higher sensitivity of the methodological approach applied in this study. Thus, more tiny subclones existing at early stages were detectable. In addition, this might also be because of a fast evolution occurred in very late samples that were only available in patients at a late stage, as those included in our study.

In our CNA analysis for a single case harboured *TP53* mutation as detected by the NGS, we showed clonal evolution as an increase in the number of involved bands of previously affected chromosomes and emergence of new chromosomal regions at time of disease relapse.

The clinically important finding in this chapter was the effect of targeted gene mutation on TFS when TFS was calculated from time of sampling (Figure 4.6), but not from time of diagnosis. This is possibly because the gene mutations might not have occurred long before the screening or may be due to some drivers required to reach a certain VAF% to exert their effects. A shorter TFS was observed for cases harbouring at least one target gene mutation. More interestingly a statistically significant association of shortest TFS with driver mutations (defined using the pre-determined VAF thresholds) was observed in this cohort. This finding further supports the influence of driver mutations in disease progression.

Taken together, these findings in this longitudinal study firstly show the necessity of deep sequencing approach to identify minor mutant subclones of prognostic potential. Secondly, not all convergent mutations play the same role in subclonal evolution. Therefore,

identification of the deleterious mutations is important for more targeted therapies which are available now. Thirdly, AID-related mutations play a role in CLL subclonal evolution, although it is unlikely to be the only factor. Fourthly, among these recurrent mutated genes, some (including *TP53*, *SF3B1*, *NOTCH1* and *BIRC3*) contribute to the subclonal evolution more profoundly than others. This will help to target these fast expanding subclones in order to maintain a balance between the subclones and preserve the clonal stability [242].

Chapter 5. Applying the ultra-deep NGS test for CLL recurrent mutations in clinical samples

5.1. Introduction and aims

As described in General Introduction, the recent availability of a number of targeted therapies for various cancers, including CLL may pave the way for personalised medicine in the future [303]. To a great extent, the success of such treatments relies on the genetic profile of the individual tumour being treated; therefore, new clinical trials to select treatment based on specific mutations found in individual patient may be underway in this hospital (region).

Having developed the fast, inexpensive and highly sensitive NGS test for multiple gene mutations in stored mononuclear cells of CLL patients, we planned to use this test as a clinical laboratory service. However, clinical samples are different from those used in the development of this test. They include fresh blood and formalin fixed paraffin embedded (FFPE) tissues. The aim of this chapter was to examine whether this test is suitable for these clinical samples, and establish a format for result reporting.

5.2. Materials and methods

5.2.1. Clinical samples

A blood sample from a newly diagnosed CLL patient and a formalin fixed paraffin embedded (FFPE) tissue sample from a patient with a history of CLL and lacrimal sac involvement were collected from the out-patient clinic of Royal Liverpool University Hospital and used in this testing. The clinical information of these cases is summarised in Table 5.1.

Table 5.1. **Overview of clinical information of the cases used in this chapter**

Cases	Age	Gender	Diagnosis	Binet stage	Therapy prior to sampling
1	50	Male	CLL	A	None
2	89	Male	CLL with lacrimal sac involvement	C	Chlorambucil and bendamustine

5.2.2. DNA extraction and NGS procedures

The blood sample from the CLL patient was processed for mononuclear cell separation and the DNA was extracted as described in Sections 2.2.1 and 2.2.2, respectively. For the lacrimal sac sample, g. DNA from FFPE tissue sample was provided from the Pathology Laboratory of Royal Liverpool Hospital. After performing the quality control checks for each g. DNA sample as mentioned in Section 2.2.3, 225 ng and 675 ng of g. DNA from the respective newly diagnosed CLL and the CLL case with lacrimal sac involvement were used for target enrichment using HaloPlex probes. The DNA samples were successfully enriched using the same approaches stated in Section 2.2.8. Later, the enriched exon DNA of the 15 genes were sequenced on an Ion 318 chip as described in Sections 2.2.11, 2.2.12 and 2.2.13. Finally, the NGS data was processed as defined in Section 2.2.14.

5.3. Results

5.3.1. DNA quality and read length of the DNA libraries

As shown in Table 5.2 and Figure 5.1, a sufficient amount of good quality of g. DNA was obtained from both the FFPE sample and the fresh peripheral blood mononuclear cells (PB MNCs). As expected, compared to the g. DNA from fresh PB MNCs and stored PB MNCs the integrity of the FFPE tissue g. DNA was low, with the range of sizes being 35 - 400 bp (also shown in Section 2.3.1.1), compared to 1000 - 10380 bp of fresh and stored MNCs. However, the optimised HaloPlex protocol enabled a successful enrichment of targets in both DNA samples within an acceptable size range: 150 - 450 bp for the FFPE and 150 - 550 bp for the

fresh PB MNCs DNAs, although the average size of the former was smaller than the latter (Figure 5.1). Of note, the starting amount of g.DNA from the FFPE tissue was 3 folds of that from the blood MNCs in order to obtain similar quantity of target DNA libraries for sequencing.

Table 5.2. **Quantities and qualities of g. DNA and DNA library size of each sample**

Samples g.DNA	Source	Quantity ng/µl	Quality OD 280/260	g. DNA size peak (bp)	DNA library size range (bp)	DNA library yield
1	FFPE tissue	97.8	1.97	< 400	150-450	2.4 ng/ µl
2	Fresh PB MNCs	65.8	1.89	>1000	150-550	8.8 ng/ µl

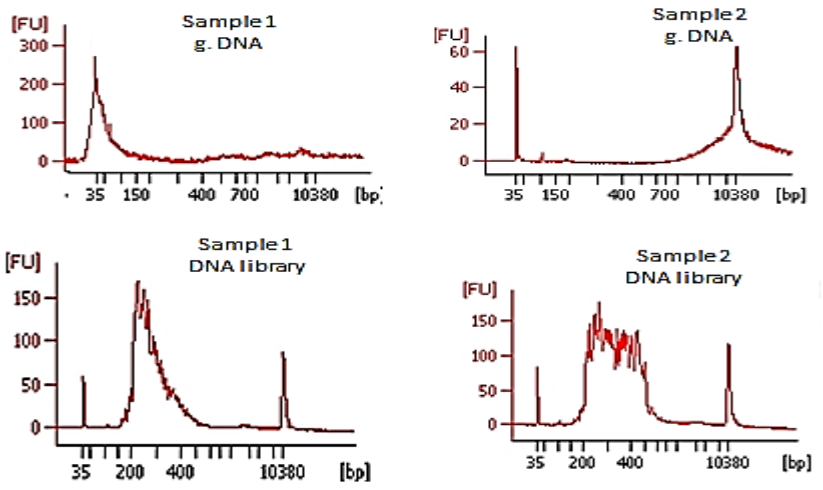


Figure 5.1. **The g. DNA integrity and DNA library peak size of the 2 clinical samples** Bioanalyser electophoreograms show size distribution (bp) of partially degraded g. DNA of a FFPE tissue (sample 1) and an intact g. DNA from the fresh blood MNCs (sample 2) and of the corresponding DNA libraries after enrichment with HaloPlex and XP bead purification

5.3.2. Sequencing quality

Satisfactory chip loading density of 77% was achieved and the sequencing data was of good quality as determined by the number of bases with AQ 20 which was 537.5 Mbp compared to the respective average values of 71% and 663 Mbp for the CLL cohort (Chapter 2, Table 2.8). However, as expected the average target DNA size (Figure 5.1), the average coverage depth and the target coverage (Table 5.3) for the blood sample was undoubtedly better than the FFPE sample, despite of the same amount of target DNA being sequenced.

Table 5.3. **Read length and quality of coverage for individual samples**

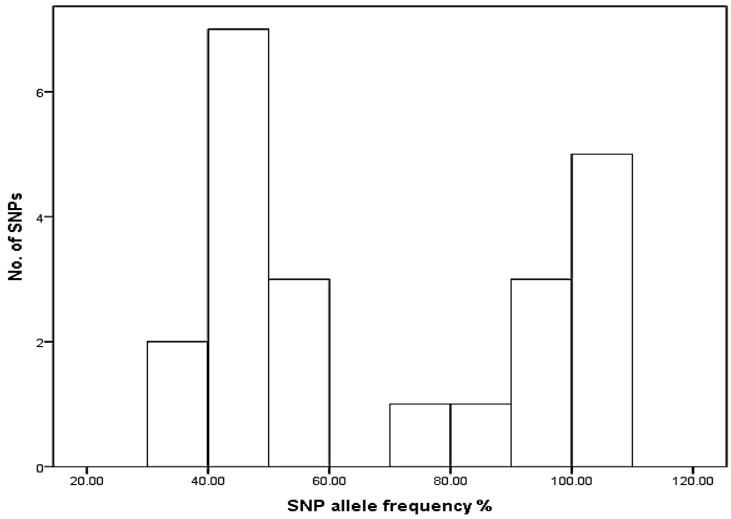
Coverage parameters	Fresh PB MNCs sample	FFPE tissue sample
Mean read length	172 bp	144 bp
Av. base coverage depth	5792 x	1110 x
Av. uniformity of coverage	95.47	89.88
Target base coverage at 1 x 20 x 100 x	 99.93% 99.76% 99.16%	 99.75% 98.24% 95.7%

5.3.3. Good test reliability as assessed by germline SNPs

We used the number of SNPs identified in each sample library to estimate target coverage. It was 0.42 and 0.46 SNPs/Kbp for the lacrimal sac and the blood sample, respectively. As compared to mean number of SNPs identified in the CLL cohort (Mean ± SD: 0.41 ± 0.04) (Section 2.3.2.1) the results were very close. As expected, the allele frequencies of most SNPs identified were close to 50% or 100%, with those in the CLL sample being closer than in the lacrimal sac sample as shown in Figure 5.2. Moreover, even a SNP located in the GC- rich region of exon 4 of *TP53* (17:7579472 G>C, p.P72R) was clearly detected in both samples, with a VAF of 95.3% and 98.8% for the lacrimal sac (FFPE tissue) sample and the CLL (fresh

blood mononuclear cell) sample, respectively. Inclusion of these controls minimised false negatives of the test for these cases.

A



B

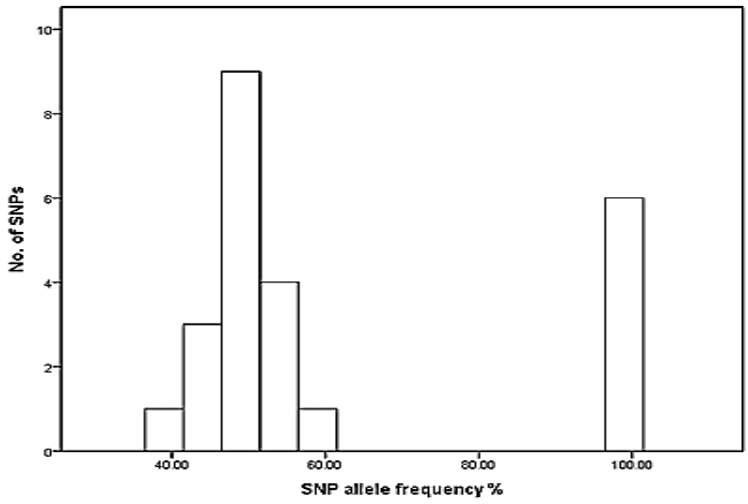


Figure 5.2. **SNP allele frequency distribution of the clinical samples**

SNPs allele frequency distribution for the lacrimal sac sample (**A**) and the CLL sample (**B**) centred around 50% and 100%

5.3.4. Gene mutation status (somatic non-synonymous variants) identified

In the total of 25 variants detected in the lacrimal sac sample, there were no somatic non-synonymous mutation, but 2 synonymous mutations with VAF of 10% and 11% in *LRP1B* (2:141259394 A>G) and *SAMHD1* (20:35579942 G>C) were identified, respectively. While, in the CLL case, a single nucleotide insertion was identified in exon 10 of *ATM* (11: 108121602-3 insertion A, VAF: 12.5%, target base coverage depth: 1998 x, the variant allele coverage: 250 x, strand bias: 0.77 and AQ: 194.6) that resulted in truncation of the protein as recorded in the report below (Figure 5.3).

5.3.5. Format of the clinical report produced

In the result report, we firstly showed quality of the test, including the average coverage depth, uniformity and limit of detection (LOD). The LOD was calculated based on a minimum of 20 mutant alleles detected for each mutation in a sample. For somatic non-synonymous mutations detected, the exact nucleotide and amino acid sequence changes were reported along with the location at corresponding chromosome and the VAF. Owing to nature of this NGS method, an absolute (100%) uniformity could not be reached. Thus, not all the target regions were evenly sequenced. Accordingly, some mutations with very low VAF might not be detectable. We therefore reported the average limit and the lowest limit of detection, as well as the average and lowest coverage depths for all target genes with and without mutations detected, and made a disclaimer to avoid any misunderstanding of negative results (Figure 5.3).

Report: CLL-NGS2015-001

The Royal Liverpool and NHS
Broadgreen University Hospitals NHS trust
Lymphoid R&D Biomarker Laboratory
Royal Liverpool University Hospital
Prescot Street, Liverpool L7 8XP
Tel: 01517064326

Mutation test of a CLL gene panel

Materials and method:

DNA extracted from blood mononuclear cells was used for enrichment of target exons using HaloPlex probes designed for 15 genes recurrently mutated in CLL. The resulting DNA was amplified by PCR and then purified with XP beads followed by sequencing on an Ion Chip 318 using the Ion Torrent PGM platform. The sequence data were processed and analysed using the Ion Reporter (V4.2) software package with the human genome sequence Hg19 being the reference. Known SNPs and synonymous mutations were filtered out in the analysis. Overall sequencing quality is summarized below:

Average depth of base coverage	Uniformity of base coverage	Average limit of detection
5792x	95.47	0.3% mutant alleles

Results:

The non-synonymous somatic genomic alteration detected are presented in the table overleaf.

Conclusion

A single nucleotide insertion in ATM resulting in protein truncation was detected in 1.5% of alleles in this sample

There are no detectable sequence alterations in the other 14 genes. However, mutations affecting a small proportion of alleles cannot be excluded due to low coverage depth in some target regions.

Genes	Target exons	Av. Coverage	Av. depth	Mutation detected	Details of the mutation identified			
					Location	c.DNA variant	Variant Protein	Mutant/total alleles
TP53	2-11	100x100.00 %	4874x	No	11:10611602	c.1410-1411insA	p.N471Kfsx16	250/1998 12.5%
ATM	2-63	100x99,16 %	5756x	Yes				
SF3B1	7-19	100x100.00 %	4984x	N0				
PCLO	2-24	20x99.77% 100x98.48 %	5882x	No				
BIRC3	2-9	100x99.37 %	6465x	No				
NOTCH1	34	20x99.87% 100x94.60 %	1674x	No				
MYD88	2-5	100x100.00 %	7163x	No				
LRP1B	34-86	100x100.00 %	6304x	No				
SAMHD1	1-16	100x100.00 %	5065x	No				
XPO1	11-16	100x100.00 %	4550x	No				
FBXW7	6-12	100x100.00 %	6147x	No				
CHD2	12-36	100x100.00 %	6929x	No				
HIST1H1E	1	100x100.00 %	5054x	No				
ZFPM2	8	100x100.00 %	6581x	No				
POT1	5-19	20x100.00 % 100x98.00 %	6130x	No				

Figure 5.3. **An example report of NGS gene panel test for CLL recurrent mutations in blood MNCs of a newly diagnosed CLL case**

5.4. Discussion

Gathering information based on gene mutations that are expected to confer resistance to a target therapeutic agent will be important for imminent deployment of effective combination therapies. Given that more genetic determinants in CLL that drive disease progression and resistance have been discovered, there exists a need to adopt a targeted NGS approach that can affordably, quickly and reliably sequence multiple genes from individual patient samples. This is because routine use of whole genome or whole exome sequencing is likely to be a few years away due to lower sensitivity and vast amount of data produced which is difficult to be managed with the available platforms and bioinformatics infrastructure in clinical laboratories. Moreover, in the WGS and WES techniques a large number of mutations can be incidentally found without recognising their significance which poses a great reporting challenge for the laboratories [304]. Therefore, we sought to implement the developed targeted NGS approach to forge the way for entering mainstream diagnostic use. As, there were no formal standards for the minimum depth of targeted sequencing we determined the sensitivity (limit of detection) based on the average coverage depth. However, the obtained depth of coverage is not always identical in all targeted regions for NGS, suggesting that some of small-size mutations with a shallow coverage may not be detectable. Therefore, target regions with minimal coverage were defined to avoid false negative results in the clinical report.

As discussed earlier, FFPE samples are invaluable both in research and clinical context. DNA library construction for NGS tends to be challenging for such samples as it has been reported that only 30% of such specimens could achieve successful enrichment [305, 306]. Here, our sample processing protocols have allowed us to robustly enrich g. DNA extracted from a FFPE sample which clearly showed degradation. We used an additional quality control step (FFPE DNA integrity analysis described in Section 2.3.1.1) by amplifying different fragment sizes of *GAPDH* before sample enrichment with HaloPlex and this provided hints that DNA fragments up to 300 bp can be successfully amplified (Figure 2.2.C). Moreover, we used greater amounts of g. DNA for the FFPE sample which was 3 folds higher than the amount used for the CLL sample. Though the amount of DNA library recovered was sufficient for sequencing, the yield of DNA library for the FFPE sample library was lower than that of the CLL sample. It is important to mention that, due to the natural flow of haematological

malignancies, the tumour specimen tends to be homogenous and thus all the subclones are equally represented [242, 307]. In contrast, a solid tumours' tissue section, including the FFPE tissue sample used in this chapter may not be representative (commonly contaminated by the surrounding normal tissue) to the whole tumour or subclones [242] and therefore additional steps are needed to evaluate and quantify histology of the specimen to establish percentage of tumour cells in the specimen based on staining in addition to examining the effect of storage time on the NGS results.

Despite that, due to the capacity of the enrichment probes in capturing short DNA fragments, this approach seems to be ideal for enrichment of FFPE derived DNA which frequently appear in degraded forms. However, the ideal number of cells or tissue sections for sufficient DNA extraction needs to be evaluated. Moreover, experiments for improving the quality of extracted DNA from FFPE tissue need to be conducted to improve the quality of the subsequent NGS results.

All in all, our results are promising and suggest that the developed NGS technique can be used to study FFPE specimens and both fresh and stored (frozen) blood specimens. Moreover, the developed reporting process was suitable for clinical application as the communicated information has recently been published in guidelines for NGS reporting in clinical settings [257, 308].

Chapter 6. General discussion

The ultimate aim of this study was to develop predictive genomic biomarkers for CLL progression and response to treatment. The objectives were firstly, to establish a gene mutation profile of samples taken at advanced stages of the disease and secondly, to study clonal evolution dynamics in order to understand inter-clonal relations and association between target gene mutations and clinical outcome.

Achievement of this goal required successful development and implementation of a highly sensitive method to monitor recurrent gene mutations in this disease. Among different NGS approaches, HaloPlex target gene enrichment and sequencing on Ion Torrent PGM were chosen for this project owing to their advantages described in the General Introduction. As both techniques were new, optimisation step by step from probe design to final reporting was essential. The work presented in chapter 2, regarding the development of the method included a series of assessments of the technique. A customised HaloPlex probe set for a panel of 15 genes reported to be recurrently mutated in CLL was designed. Together with carefully adjusted other conditions, this allowed good cost efficiency and a high test sensitivity. Thus, the cost of consumables for each test was only £265, and the average detection limit was as low as 1% VAF, if 4 DNA libraries were sequenced on each chip with an expected coverage depth 1846 x of 135.42 Kbp of sequenceable region.

In our study, the average target coverage depth (2252 x) obtained was actually bigger than expected for the sequenceable regions, this was because the expected coverage depth was calculated assuming an even distribution of coverage across the target sequences and their adjacent off-target regions. However, the coverage for target regions was always much bigger than that for the off-target, due to the probe design.

On the other hand, even in these target regions, the coverage was not absolutely even. This is an issue for NGS, which has been also reported by others [309], and was particularly important when certain regions with low coverage depth and no mutation detected. In this respect, false negative results might be reported. We were therefore very cautious in the report by including the average coverage depth of each target gene to avoid a misinterpretation of negativity to be wild type. This is still an issue for future discussion in

the clinical application of this test. Another finding in the initial phase of our study was the missing target regions with GC-rich sequences, including the 75 bp in exon 4 of *TP53*. Though small, these missing regions were not negligible and therefore were covered by a boost of probe overlay in our subsequent HaloPlex probe design which resulted in dramatic increase of the coverage depth of this GC rich target region. Thus, this strategy is useful for any other GC rich targets with low amplicons to augment the coverage depth (Figure 2.5).

Considering the wet lab procedures, good quality DNA is essential for NGS. Moreover, successful library amplification and purification are key steps of HaloPlex target enrichment. Initially, we used the manufacture's protocol, as shown in Figure 2.3, for the DNA digestion step; the fragments were larger than expected. Therefore, a longer DNA incubation time with the enzymes was necessary for appropriate DNA digestion. Furthermore, no DNA library was initially recovered following the manufacturer's instructions. In our experiments, elongating the duration of XP bead-library incubation could overcome the problem and sufficient amounts of DNA libraries were obtained (Figure 2.7). In addition, sufficient purity was achieved using double purification by XP-beads (Figure 2.8) which later maximised the chip capacity usage for deeper sequencing. We also reduced the amounts of probes without any effect on DNA library amplification. This strategy improved the efficiency of purification and saved as much as 16% in the amount of the probes used (Section 2.3.1.5).

Data analysis is as equally important for this NGS test. We used the manufacturer-recommended analysis pipeline in the analytical process. However, the defaulted high and low stringency settings of the TVC plugin initially used were not designed for the purpose of this test for these recurrent mutations in CLL. Based on known VAF% of 5 different mutations scattered in *TP53*, we were able to test reliability of these defaulted settings. As shown in Table 2.11, false negative and false positive variant calls were found. Our optimised variant calling plugin with parameter settings that had yet acceptable confidence scores improved both sensitivity and precision of the test to 100%. In subsequent As-PCR experiments using serial diluted DNA, we confirmed the fidelity of these mutations identified by the PGM and verified the high sensitivity of the NGS test we developed (Section 2.3.3). Besides, we tested the reproducibility of the test by duplicate sequencing runs of multiple CLL samples. As expected, a high agreement between the tests for both the variant detection and VAF% in all samples were achieved (Figure 2.13). More convincingly, the

alternative validation of *ATM* mutational status in 7 samples by the University of Birmingham and the high agreement of the sequencing results of 5 CLL samples screened by ERIC for *TP53* mutational status (multicentre double blinded study) confirmed the reliability of our NGS test (Section 2.3.3.4). The conclusions that can be drawn from Chapter 2 are that all the optimised steps including the data analysis generated robust results. Thus, this test can efficaciously be implemented to achieve other immediate (Chapter 3, 4 and 5) and future aims of this study.

With the reliable NGS test established, we next implemented this approach in Chapter 3 to study the mutation profiles in 32 CLL patients with progressive and/or chemo-resistant disease along with another patient with an indolent disease. We further documented the high quality of data using SNP pattern as a genomic imprint [250] and Ti/Tv ratio to the expected levels reported by others [254]. These guaranteed the reliability of the data we had prior to downstream analysis. Although it was easy to distinguish non-synonymous from synonymous mutations, we were careful in identifying somatic alterations as there were no germline DNA controls. We determined somatic alterations by comparing the variants to records in dbSNP (for germline variants) and COSMIC (for somatic variants) databases. We also used the VAF of 2-40% or 60-90%, and the changeable VAF when serial samples were sequenced as additional criteria for somatic mutations if they were not previously reported.

Of note, the gene panel was selected at the beginning of this study and therefore did not include recurrent mutated genes identified afterwards [278, 280]. Our results might not reflect the whole picture of the mutation profiles in this cohort of CLL. Nevertheless, we identified 79 somatic non-synonymous mutations in 28 cases. We could validate 89.87% of all these identified variants by various methods including Sanger sequencing, repeated NGS, As-PCR and the subsequent longitudinal study. Therefore, by alternative methods including Sanger sequencing and As-PCR, 16 point mutations and 10 indels with VAF ranging between 5.0% to 98% in genes recurrently mutated in our cohort, including *TP53*, *ATM*, *SF3B1*, *PCLO* and *NOTCH1* were verified (Figure 3.8).

Although we have not targeted the non-coding region of *NOTCH1* [278] and other genes that have only recently been identified in CLL including *FAT1* [279], *NFKBIE* [310], *BTK* [29] and *RPS15* [280], in this chapter we could identify mutations in 12 of the 15 selected genes in the

28 patients at the latest stages of the disease (Table 3.6). In total, 28%, 31%, 34% and 25% of cases harboured mutations in *TP53*, *ATM*, *SF3B1* and *PCLO*, respectively. They were much higher than the frequency reported in other studies [95, 135, 186, 281]. This might reflect a difference of patient components in our cohort, as all were at advanced disease stages. We found a dispersal distribution of mutations in *TP53* including a mutation in exon 11, and a mutation outside the Heat repeat of *SF3B1* (exon 11) which are often not included in most of studies [158, 311]. This suggested that mutations outside hot regions exist in CLL and their biological and clinical impact should be studied further. Moreover, in *ATM* which is a large target gene, we found a widespread distribution of identified mutations in this cohort. In general, 37.97% of the mutations in all genes were between 2% and 20% VAF. Although those mutations are difficult to be detected with other methodologies, including conventional Sanger sequencing, it is difficult to ensure that mutations below 5% VAF are true mutations because of the absence of bio-informatic support in this study as well as lack of a reliable methodology to validate these mutations.

We showed the heterogeneous subclonal composition between the cases. Not surprisingly, all the *TP53* mutated cases and 80% of the *ATM* mutated cases had at least another gene mutation. We assumed that, in any patient, mutations with a higher VAF might have occurred earlier than mutations with a lower VAF or might have had more rapid clonal evolution and worse prognosis. Based on the VAF, we explored the dominance of these mutations. We revealed that *ATM* and *TP53* mutations were dominant in 73.7% of the cases (Section 3.3.9), compared to 37.8% of other gene mutations, this was significantly different ($P = 0.011$). We also showed a borderline association of gene mutations and chemotherapy (Section 3.3.8), an increased number of mutation events or mutated genes were documented in cases who received chemotherapy prior to sampling with p values close to 0.05. A larger number of samples are required to draw definitive conclusion about the existing relationships. Furthermore, we revealed a significant association between the number of chemotherapy cycles and multiplicity of gene mutations ($P = 0.036$). Considering the study of copy number changes in the 13 CLL cases, we identified the recurrent changes for CLL, including 11q-, 13q-, 17p- and chromosome 12+ and other alterations affected chromosome 2, 4, 6, 7, 8, 9, 10 and 20 (Table 3.8). Our analysis of genes involved in these chromosomal aberrations revealed an enrichment of some important CLL and cancer genes

which functioned in DNA damage/repair, cellular proliferation and apoptosis pathways. We reported the involvement of chromosome 17p, 11q and 2q in patients with *TP53*, *ATM*, *BIRC3* and *SF3B1* mutations. We also identified the copy neutral loss of heterozygosity affected 9q in a case who harboured two truncating *NOTCH1* mutations. Moreover, we showed a statistically significant association of *TP53* and *ATM* mutations with increased number of mutation and CNA events (P = 0.026).

Conclusions that can be drawn from this chapter include that the dispersal distribution of mutations in important genes and the identification of a high proportion of mutations with low VAF, further stress the need of a high throughput and more sensitive methods for both research and clinical use. In addition, the high incidence of *TP53* and *ATM* mutations and their dominance over the other mutations suggests that a defective DNA damage response pathway plays an important role in both the acquisition of additional mutations and in the development of an aggressive disease phenotype. Moreover, the association of chemotherapy and number of gene mutations suggests the possible role of treatment in the induction or selection of mutations.

Paired serial samples were used for the study of subclonal evolution in Chapter 4. We firstly performed strict quality control checks to exclude any samples lacking matched SNP fingerprint (Section 4.3.1). This guaranteed the reliability of our analysis. Thus, 33 tumour samples from 23 CLL cases with a perfect SNP match were analysed. In the analysis of nucleotide changes in the context of tri-nucleotides, firstly, we showed the predominance of C:G>T:A changes as also found by others [288, 290]. We identified an increase in C>T changes after therapy in 14 CLL cases (Figure 4.1). By examining the mutation signature at earliest stages and at latest stages for the 21 serial samples with single nucleotide substitutions, we were able to divide all of these mutations in to AID-, ageing- and other factors-related mutations. Examination of the occurrence of each mutation type in patients who had available *IGHV* mutation data revealed that UM-*IGHV* cases had higher incidence of ageing associated mutations. This was consistent with findings by others [290] and suggested a proliferation history in UM-*IGHV* CLL.

Exploring the role of each signature in clonal evolution, we statistically confirmed the role of AID in ongoing mutations in its off-target genes, because AID-related mutations were highly

associated with the subclonal evolution. This was a good evidence supported a notion made by Kasar et al based on a single time point WGS study that AID contributes to ongoing mutations in genes other than IGV in CLL [290]. However, AID seemed to not be the only driving force in such CLL clonal evolution as we found other factors that were also related with this process according to our study. Although further studies are required for understanding what exactly those factors are, APOBEC3 may be one of them. As suggested by others, distinguishing APOBEC3 signature requires sequencing of c.DNA to identify transcription bias characteristic of this enzyme produced by transcription-coupled DNA repair [312].

Considering the clonal evolution in general, we found that 92.5% of the mutations detected at the latest time points were also identified at the earliest time points, indicating a dominance of the linear evolution over the branch evolution. Of the 65 mutations identified at the earliest time points in the 23 CLL cases, 60% of them underwent clonal evolution, 18% were stable and 22% underwent clonal devolution. This phenomenon has also been found by others [128]. In our cohort, only 5 (7.69%) mutations with VAF ranged 1% - 4% identified at early stages became undetectable at the latest time points. This was more likely due to the VAF reduction of those non-dominant subclones below the limit of detection of this test, with increase of the dominant subclones during disease progression. Moreover, convergent evolution and clonal dominance of deleterious mutations occurred in 2 patients with *ATM* and *TP53* mutations, respectively. Further analysis revealed the prioritisation of deleterious mutations in convergent clonal evolutions in each of the cases. However, in this study we could not calculate the size of the mutant clones as performed by others [198,302]. This was due to incomplete information about leukocytes and lymphocyte counts. Therefore, “devolution” was the relative reduction of VAF, not necessarily meaning shrinkage of the subclonal size.

Moreover, in the longitudinal CNA study, we reported the increase in the number of regions affected by the CNA as well as expanding in size of previously affected regions. There was an overt chromothripsis-like event (defined by ≥ 5 changes in copy number state affecting a single chromosome in a single event) that affected chromosome 6 of CLL-7 (Table 4.7). However, definitive identification of this event needs to rule-out gradual accumulation of the copy number changes. In this CLL patient, the number of the bands affected by the CNA

in both samples studied fulfilled this criterion, although we identified the addition of another band in the latest sample, yet a gradual accumulation of all the affected bands in the earlier sample studied need to be excluded. Therefore, additional earlier time points were required to rule out the possibility of chromothripsis by finding that the memory genetic lesion existed with the same composition. Moreover, for a conclusive identification of chromothripsis we also need to have information about the occurrence of the CNA changes on the same homolog. This can only be obtained through long read paired end WGS [313]. Chromothripsis is a rare incident in the development and progression of CLL. If it occurs, it is a massive genetic change in a single catastrophic event rather than the commonly believed gradual accumulation of genetic lesions in tumour progression [314].

Considering the clinical impact of mutations, clonal and subclonal *TP53* mutations tend to have similar clinical consequences [138]. Whether this is true for other recurrently mutated genes deemed to be potential drivers (*NOTCH1*, *SF3B1*, *BIRC3* and more recently, *SAMHD1*) has not been well-established and still remained controversial [315]. To clarify this matter, as in our cohort, based on the observation that various gene mutations within the same tumour sample showed a different speed of evolution, we calculated the subclonal VAF doubling time for each target gene mutation regardless of their co-associated mutations. Our results showed that the above potential driver mutations had shorter VAF doubling time compared to other targeted genes (Figure 4.4). In agreement with the classification of these driver mutations under high risk category [316], our results confirmed their association with rapid clonal evolution as compared to *ATM* which has been put under intermediate risk category. This observation further supports the notion that these high-risk gene mutations are associated with unfavourable CLL outcome [317].

In the analysis for clinical impact of the genes mutation on TFS and OS, we revealed that patients possessing at least a target gene mutation had a shorter TFS (calculated from the time of sampling time to the 1st treatment) compared to patients with no mutations. Moreover, within the mutated group, patients who harboured at least a driver gene mutation (including *TP53*, *SF3B1*, *NOTCH1*, *BIRC3* and *SAMHD1*) with a certain subclonal size as previously set [293] were significantly associated with a shorter TFS (Figure 4.6).

Henceforth, given the relatively small number of cases in the studied cohort in general and among them small number of cases harbouring different gene mutations and the variability in the VAF at presentation, most of our findings are preliminary. It would be worth performing larger-scale (increase the number of patients) ultra-deep NGS studies in the future in order to analyse clonal/subclonal mutations with an effective bio-informatics support to draw solid conclusions regarding the relative impact of each type of mutation as solitary abnormality or in conjunction with the presence of other genetic events.

Taken together, the results in Chapter 4 revealed the importance of AID in CLL evolution; therefore it might be a candidate gene for future studies on CLL biology and treatment. Moreover, the linearity of clonal evolution predominantly presented in our cohort suggested the requirement for this highly sensitive NGS test to detect the small mutant subclones at as early as possible, so that the genetic composition at the latest stages can be predictable from the early stages. However, the roles of deleterious mutations in convergent evolution suggest that not all the mutations within the same target gene have the same biological and clinical impacts. Therefore, how the subsequently dominant convergent mutations can be predicted at an early stage in CLL requires further study, likely in a cooperation of multiple centres.

In our study of clinical implementation of this test (Chapter 5), we clearly showed the effect of degraded DNA extracted from formalin fixed-paraffin embedded tissue on quality of NGS (Figure 5.1). Despite of that, the results generated by us were promising and HaloPlex might be suitable for target gene enrichment of FFPE samples. However optimisation for DNA extraction from FFPE to improve quality of the resulting DNA is a key step for success in this NGS test. Moreover, for the DNA sample extracted from fresh mononuclear cell, a superior performance of HaloPlex and Ion Torrent NGS has been shown as we demonstrated the high quality of data and the extra-ordinary high coverage depth for that sample (Table 5.3). Nevertheless, to our concern, the uniformity of coverage should be always checked and regions with low coverage should be reported along with the limit of detection as we did in our result reporting (Figure 5.3) in order to avoid misunderstanding negative results.

In conclusion, we have developed a sensitive and reliable in-house-NGS method to screen recurrent mutations in CLL. Moreover, the most frequently mutated genes included *SF3B1*,

ATM, *TP53* and *PCLO*. Moreover, ATM-p53 pathway was the most common mutated pathway at late stages of disease. The association of this pathway with clonal evolution emphasises its importance in CLL progression. However, *SF3B1*, *NOTCH1*, *SAMHD1*, *CHD2*, *PCLO* and *LRP1B* can also be the dominant mutated gene. Importantly, due to a linear evolution pattern, most of the dominant mutations at late stages were detectable at early stages prior to treatment, and the therapeutic intervention contributed to the clonal selection. Therefore, it is important to detect these mutations as early as possible to help guide clinician's treatment choice for individual patients. Therefore, a highly sensitive ultra-deep sequencing as well as bio-informatics support are required to meet this clinical demand. In our study, the documented convergent clonal evolution has suggested that deleterious mutations are predominant in this phenomenon, but further studies are required to validate this. Moreover, the relationship between driver mutations and subclonal evolution need confirmation by further research work. For the study of clinical impact, we can no longer merely consider the presence or absence of mutations, it will be vital to know what proportion of cells harbour a mutation and what the functional consequence of mutation is. This will influence clinical decision making such as predicting outcome and choosing the right therapy. This information might also be important to optimize the dose of therapy. Along these lines, it might be imperative to address the issue of subclonal mutations in extended cohorts of homogeneously treated patients particularly in this new era of targeted therapy such as using inhibitors of BCR signalling and the newly developed pathway specific small molecules inhibitors. Therefore, target genes that have been recently identified in CLL, for example, *RPS15* [280], *BTK* and *PLCG* [318], may need to be included in the gene panel in future.

Chapter 7. Appendices

7.1. HaloPlex probe design information

Table 7.1. **Chromosomal coordinates used to design HaloPlex probes for this study**

Genomic coordinate	Gene Name	Exon No.	Ensembl Exon ID
chr11:102195193-102196093	*BIRC3*	2	ENSE00001256551
chr11:102196197-102196296	*BIRC3*	3	ENSE00002474565
chr11:102198783-102198861	*BIRC3*	4	ENSE00000745287
chr11:102199628-102199676	*BIRC3*	5	ENSE00000795265
chr11:102201730-102201972	*BIRC3*	6	ENSE00000795266
chr11:102206697-102206951	*BIRC3*	7	ENSE00000745300
chr11:102207491-102207532	*BIRC3*	8	ENSE00002467460
chr11:102207640-102207880	*BIRC3*	9	ENSE00002162318
chr11:108098322-108098423	*ATM*	2	ENSE00003535574
chr11:108098503-108098615	*ATM*	3	ENSE00003590622
chr11:108099905-108100050	*ATM*	4	ENSE00003614955
chr11:108106397-108106561	*ATM*	5	ENSE00001667088
chr11:108114680-108114845	*ATM*	6	ENSE00001670710
chr11:108115515-108115753	*ATM*	7	ENSE00001739598
chr11:108117691-108117854	*ATM*	8	ENSE00001617295
chr11:108119660-108119829	*ATM*	9	ENSE00001652815
chr11:108121428-108121799	*ATM*	10	ENSE00001658306
chr11:108122564-108122758	*ATM*	11	ENSE00001638731
chr11:108123544-108123639	*ATM*	12	ENSE00001655406
chr11:108124541-108124766	*ATM*	13	ENSE00001592116
chr11:108126942-108127067	*ATM*	14	ENSE00001774900
chr11:108128208-108128333	*ATM*	15	ENSE00001723835
chr11:108129713-108129802	*ATM*	16	ENSE00001769509
chr11:108137898-108138069	*ATM*	17	ENSE00001719378
chr11:108139137-108139336	*ATM*	18	ENSE00003595770
chr11:108141791-108141873	*ATM*	19	ENSE00001591923
chr11:108141978-108142133	*ATM*	20	ENSE00001764296
chr11:108143259-108143334	*ATM*	21	ENSE00003491738
chr11:108143449-108143579	*ATM*	22	ENSE00003484127
chr11:108150218-108150335	*ATM*	23	ENSE00003605753
chr11:108151722-108151895	*ATM*	24	ENSE00001618543
chr11:108153437-108153606	*ATM*	25	ENSE00001761338
chr11:108154954-108155200	*ATM*	26	ENSE00001649610
chr11:108158327-108158442	*ATM*	27	ENSE00001728966
chr11:108159704-108159830	*ATM*	28	ENSE00003469194
chr11:108160329-108160528	*ATM*	29	ENSE00003581035
chr11:108163346-108163520	*ATM*	30	ENSE00003598677
chr11:108164040-108164204	*ATM*	31	ENSE00003529497
chr11:108165654-108165786	*ATM*	32	ENSE00003559264
chr11:108168014-108168109	*ATM*	33	ENSE00003480406
chr11:108170441-108170612	*ATM*	34	ENSE00003565584
chr11:108172375-108172516	*ATM*	35	ENSE00003590115
chr11:108173580-108173756	*ATM*	36	ENSE00003508061
chr11:108175402-108175579	*ATM*	37	ENSE00003472092

Table 7.1. (continued)

Genomic coordinate	Gene Name	Exon No.	Ensembl Exon ID
chr11:108178624-108178711	*ATM*	38	ENSE00003472517
chr11:108180887-108181042	*ATM*	39	ENSE00003591034
chr11:108183138-108183225	*ATM*	40	ENSE00003458427
chr11:108186550-108186638	*ATM*	41	ENSE00003614552
chr11:108186738-108186840	*ATM*	42	ENSE00003552391
chr11:108188100-108188248	*ATM*	43	ENSE00003522577
chr11:108190681-108190785	*ATM*	44	ENSE00003688819
chr11:108192028-108192147	*ATM*	45	ENSE00003648389
chr11:108196037-108196271	*ATM*	46	ENSE00003542516
chr11:108196785-108196952	*ATM*	47	ENSE00003580001
chr11:108198372-108198485	*ATM*	48	ENSE00003479049
chr11:108199748-108199965	*ATM*	49	ENSE00003599084
chr11:108200941-108201148	*ATM*	50	ENSE00003659411
chr11:108202171-108202284	*ATM*	51	ENSE00003596787
chr11:108202606-108202764	*ATM*	52	ENSE00003571052
chr11:108203489-108203627	*ATM*	53	ENSE00003483502
chr11:108204613-108204695	*ATM*	54	ENSE00003573212
chr11:108205696-108205836	*ATM*	55	ENSE00003560896
chr11:108206572-108206688	*ATM*	56	ENSE00003666486
chr11:108213949-108214098	*ATM*	57	ENSE00003611442
chr11:108216470-108216635	*ATM*	58	ENSE00003671649
chr11:108218006-108218092	*ATM*	59	ENSE00003588344
chr11:108224493-108224607	*ATM*	60	ENSE00003609743
chr11:108225538-108225601	*ATM*	61	ENSE00003638878
chr11:108235809-108235945	*ATM*	62	ENSE00003502604
chr11:108236052-108236236	*ATM*	63	ENSE00001331356
chr15:93489268-93489446	*CHD2*	12	ENSE00003593607
chr15:93492182-93492306	*CHD2*	13	ENSE00001255168
chr15:93496587-93496803	*CHD2*	14	ENSE00003481778
chr15:93498653-93498742	*CHD2*	15	ENSE00003587849
chr15:93499689-93499879	*CHD2*	16	ENSE00003503986
chr15:93510555-93510743	*CHD2*	17	ENSE00001255136
chr15:93514995-93515157	*CHD2*	18	ENSE00001255131
chr15:93515495-93515647	*CHD2*	19	ENSE00003659253
chr15:93518109-93518180	*CHD2*	20	ENSE00003535318
chr15:93521464-93521613	*CHD2*	21	ENSE00002434605
chr15:93522365-93522513	*CHD2*	22	ENSE00001097892
chr15:93524045-93524141	*CHD2*	23	ENSE00001097780
chr15:93524595-93524687	*CHD2*	24	ENSE00001097863
chr15:93527560-93527730	*CHD2*	25	ENSE00001097839
chr15:93528728-93528903	*CHD2*	26	ENSE00001097827
chr15:93534706-93534747	*CHD2*	27	ENSE00001097770
chr15:93536089-93536228	*CHD2*	28	ENSE00001097848
chr15:93540187-93540325	*CHD2*	29	ENSE00001097738

Table 7.1. (continued)

Genomic coordinate	Gene Name	Exon No.	Ensembl Exon ID
chr15:93540483-93540633	*CHD2*	30	ENSE00002469647
chr15:93541729-93541851	*CHD2*	31	ENSE00003503542
chr15:93543742-93543870	*CHD2*	32	ENSE00003673545
chr15:93545407-93545547	*CHD2*	33	ENSE00003587375
chr15:93547847-93547981	*CHD2*	34	ENSE00001314469
chr15:93552375-93552553	*CHD2*	35	ENSE00001097753
chr15:93555575-93555674	*CHD2*	36	ENSE00001179528
chr17:7572926-7573008	*TP53*	11	ENSE00003605891
chr17:7573927-7574033	*TP53*	10	ENSE00003545950
chr17:7576853-7576926	*TP53*	9	ENSE00003636029
chr17:7577019-7577155	*TP53*	8	ENSE00003586720
chr17:7577499-7577608	*TP53*	7	ENSE00003504863
chr17:7578177-7578289	*TP53*	6	ENSE00003462942
chr17:7578371-7578554	*TP53*	5	ENSE00003518480
chr17:7579312-7579590	*TP53*	4	ENSE00003625790
chr17:7579700-7579721	*TP53*	3	ENSE00002419584
chr17:7579839-7579940	*TP53*	2	ENSE00002667911
chr20:35521334-35521469	*SAMHD1*	16	ENSE00000800455
chr20:35526225-35526362	*SAMHD1*	15	ENSE00000661812
chr20:35526843-35526947	*SAMHD1*	14	ENSE00000661813
chr20:35532560-35532652	*SAMHD1*	13	ENSE00003692177
chr20:35533767-35533906	*SAMHD1*	12	ENSE00000661815
chr20:35539621-35539736	*SAMHD1*	11	ENSE00001622813
chr20:35540864-35540955	*SAMHD1*	10	ENSE00001711142
chr20:35545125-35545233	*SAMHD1*	9	ENSE00001793017
chr20:35545352-35545452	*SAMHD1*	8	ENSE00001667406
chr20:35547776-35547922	*SAMHD1*	7	ENSE00001752622
chr20:35555585-35555655	*SAMHD1*	6	ENSE00003598833
chr20:35559163-35559278	*SAMHD1*	5	ENSE00003535434
chr20:35563432-35563592	*SAMHD1*	4	ENSE00003546823
chr20:35569442-35569514	*SAMHD1*	3	ENSE00003462531
chr20:35575141-35575207	*SAMHD1*	2	ENSE00003619308
chr20:35579839-35580079	*SAMHD1*	1	ENSE00001898623
chr2:141027811-141027915	*LRP1B*	86	ENSE00001132056
chr2:141031993-141032167	*LRP1B*	85	ENSE00001132065
chr2:141055377-141055538	*LRP1B*	84	ENSE00001153896
chr2:141072504-141072668	*LRP1B*	83	ENSE00001182326
chr2:141079532-141079657	*LRP1B*	82	ENSE00001153862
chr2:141081462-141081635	*LRP1B*	81	ENSE00001131995

Table 7.1. (continued)

Genomic coordinate	Gene Name	Exon No.	Ensembl Exon ID
chr2:141083331-141083447	*LRP1B*	80	ENSE00001131999
chr2:141092022-141092128	*LRP1B*	79	ENSE00001132004
chr2:141093184-141093407	*LRP1B*	78	ENSE00001132013
chr2:141108366-141108607	*LRP1B*	77	ENSE00001132019
chr2:141110522-141110641	*LRP1B*	76	ENSE00001132026
chr2:141113911-141114045	*LRP1B*	75	ENSE00001132029
chr2:141115548-141115685	*LRP1B*	74	ENSE00001132034
chr2:141116390-141116515	*LRP1B*	73	ENSE00001182394
chr2:141122230-141122352	*LRP1B*	72	ENSE00001182403
chr2:141128279-141128411	*LRP1B*	71	ENSE00001328871
chr2:141128748-141128854	*LRP1B*	70	ENSE00001182417
chr2:141130577-141130706	*LRP1B*	69	ENSE00001182423
chr2:141135749-141135855	*LRP1B*	68	ENSE00001182430
chr2:141143462-141143578	*LRP1B*	67	ENSE00001182437
chr2:141200073-141200192	*LRP1B*	66	ENSE00001153839
chr2:141201899-141202018	*LRP1B*	65	ENSE00001131958
chr2:141202132-141202248	*LRP1B*	64	ENSE00001131968
chr2:141208137-141208230	*LRP1B*	63	ENSE00001131978
chr2:141214024-141214172	*LRP1B*	62	ENSE00001131982
chr2:141215032-141215220	*LRP1B*	61	ENSE00001131988
chr2:141232707-141232906	*LRP1B*	60	ENSE00001153849
chr2:141242912-141243093	*LRP1B*	59	ENSE00001153824
chr2:141245186-141245308	*LRP1B*	58	ENSE00001131904
chr2:141250177-141250262	*LRP1B*	57	ENSE00001131912
chr2:141253134-141253317	*LRP1B*	56	ENSE00001131919
chr2:141259256-141259443	*LRP1B*	55	ENSE00001131923
chr2:141260532-141260672	*LRP1B*	54	ENSE00001131930
chr2:141264365-141264487	*LRP1B*	53	ENSE00001131937
chr2:141267497-141267625	*LRP1B*	52	ENSE00001131947
chr2:141272222-141272341	*LRP1B*	51	ENSE00001153830
chr2:141274458-141274580	*LRP1B*	50	ENSE00001182504
chr2:141283413-141283562	*LRP1B*	49	ENSE00001182509
chr2:141283806-141283919	*LRP1B*	48	ENSE00001182513
chr2:141291590-141291709	*LRP1B*	47	ENSE00001153804
chr2:141294150-141294278	*LRP1B*	46	ENSE00001131890
chr2:141298542-141298667	*LRP1B*	45	ENSE00001131896
chr2:141299348-141299540	*LRP1B*	44	ENSE00001153813
chr2:141356200-141356404	*LRP1B*	43	ENSE00001182529
chr2:141359019-141359208	*LRP1B*	42	ENSE00001182536

Table 7.1. (continued)

Genomic coordinate	Gene Name	Exon No.	Ensembl Exon ID
chr2:141457819-141458190	*LRP1B*	41	ENSE00001316880
chr2:141459290-141459414	*LRP1B*	40	ENSE00001131885
chr2:141459710-141459861	*LRP1B*	39	ENSE00001182554
chr2:141459996-141460122	*LRP1B*	38	ENSE00001182559
chr2:141473542-141473671	*LRP1B*	37	ENSE00001182564
chr2:141474251-141474385	*LRP1B*	36	ENSE00001182569
chr2:141526782-141526913	*LRP1B*	35	ENSE00001182573
chr2:141528450-141528575	*LRP1B*	34	ENSE00001182578
chr2:198264976-198265158	*SF3B1*	19	ENSE00000964872
chr2:198265439-198265660	*SF3B1*	18	ENSE00000964871
chr2:198266124-198266249	*SF3B1*	17	ENSE00000964870
chr2:198266466-198266612	*SF3B1*	16	ENSE00000964869
chr2:198266709-198266854	*SF3B1*	15	ENSE00000964868
chr2:198267280-198267550	*SF3B1*	14	ENSE00000964867
chr2:198267673-198267759	*SF3B1*	13	ENSE00000964866
chr2:198268309-198268488	*SF3B1*	12	ENSE00000964865
chr2:198269800-198269901	*SF3B1*	11	ENSE00000964864
chr2:198269999-198270196	*SF3B1*	10	ENSE00000964863
chr2:198272722-198272843	*SF3B1*	9	ENSE00000964862
chr2:198273093-198273305	*SF3B1*	8	ENSE00000964861
chr2:198274494-198274731	*SF3B1*	7	ENSE00000964860
chr2:61719170-61719333	*XPO1*	16	ENSE00003613775
chr2:61719460-61719616	*XPO1*	15	ENSE00003675145
chr2:61719702-61719883	*XPO1*	14	ENSE00003464720
chr2:61720050-61720188	*XPO1*	13	ENSE00003631255
chr2:61721029-61721226	*XPO1*	12	ENSE00000757711
chr2:61722590-61722748	*XPO1*	11	ENSE00003556668
chr3:38181355-38181489	*MYD88*	2	ENSE00003582265
chr3:38181879-38182059	*MYD88*	3	ENSE00001744479
chr3:38182248-38182339	*MYD88*	4	ENSE00003552693
chr3:38182623-38182783	*MYD88*	5	ENSE00003619131
chr4:153244031-153244301	*FBXW7*	12	ENSE00001431204
chr4:153245336-153245546	*FBXW7*	11	ENSE00000821072
chr4:153247158-153247383	*FBXW7*	10	ENSE00000821073
chr4:153249360-153249541	*FBXW7*	9	ENSE00003478818
chr4:153250824-153250937	*FBXW7*	8	ENSE00003629197
chr4:153251884-153252020	*FBXW7*	7	ENSE00003495121
chr4:153253748-153253871	*FBXW7*	6	ENSE00000821077
chr4:153258954-153259088	*FBXW7*	5	ENSE00001002596
chr6:26156619-26157280	*HIST1H1E*	1	ENSE00001173828
chr7:124464008-124464128	*POT1*	19	ENSE00001863642
chr7:124465306-124465411	*POT1*	18	ENSE00002488410

Table 7.1. (continued)

Genomic coordinate	Gene Name	Exon No.	Ensembl Exon ID
chr7:124467268-124467359	*POT1*	17	ENSE00002495808
chr7:124469308-124469396	*POT1*	16	ENSE00002463467
chr7:124475333-124475468	*POT1*	15	ENSE00002489137
chr7:124481027-124481232	*POT1*	14	ENSE00002489750
chr7:124482861-124483017	*POT1*	13	ENSE00002511886
chr7:124486996-124487052	*POT1*	12	ENSE00002441395
chr7:124491926-124492005	*POT1*	11	ENSE00001136130
chr7:124493026-124493192	*POT1*	10	ENSE00001136133
chr7:124499011-124499166	*POT1*	9	ENSE00001136138
chr7:124503404-124503694	*POT1*	8	ENSE00003688811
chr7:124510965-124511095	*POT1*	7	ENSE00001785664
chr7:124532320-124532434	*POT1*	6	ENSE00003596271
chr7:124537219-124537266	*POT1*	5	ENSE00003488402
chr7:82389955-82390100	*PCLO*	24	ENSE00002431590
chr7:82390675-82390809	*PCLO*	23	ENSE00002462574
chr7:82430834-82430907	*PCLO*	22	ENSE00002457830
chr7:82435004-82435145	*PCLO*	21	ENSE00003509997
chr7:82451811-82452005	*PCLO*	20	ENSE00002527473
chr7:82453552-82453732	*PCLO*	19	ENSE00003647047
chr7:82455905-82455976	*PCLO*	18	ENSE00003644760
chr7:82457189-82457282	*PCLO*	17	ENSE00003501634
chr7:82464983-82465009	*PCLO*	16	ENSE00002459797
chr7:82467534-82467658	*PCLO*	15	ENSE00003621206
chr7:82470775-82470825	*PCLO*	14	ENSE00003532351
chr7:82474587-82474801	*PCLO*	13	ENSE00003668844
chr7:82475883-82475950	*PCLO*	12	ENSE00003573676
chr7:82476455-82476563	*PCLO*	11	ENSE00003581347
chr7:82508653-82508778	*PCLO*	10	ENSE00003503371
chr7:82531967-82532057	*PCLO*	9	ENSE00002490267
chr7:82538193-82538329	*PCLO*	8	ENSE00002529510
chr7:82544002-82546189	*PCLO*	7	ENSE00002432640
chr7:82578792-82580806	*PCLO*	6	ENSE00002709222
chr7:82581172-82586251	*PCLO*	5	ENSE00001715762
chr7:82595087-82595803	*PCLO*	4	ENSE00001646940
chr7:82763566-82764972	*PCLO*	3	ENSE00001774996
chr7:82784064-82785708	*PCLO*	2	ENSE00001622512
chr8:106813275-106815770	*ZFPM2*	8	ENSE00002135760
chr9:139390510-139392010	*NOTCH1*	34	ENSE00001247584

7.2. Recipes of routinely used solutions

Table 7.2.**Contents of 50 mM NaOH solution**

Components	Amount
Sodium hydroxide (NaOH) (40 g/L)	0.2 g
De-ionised nuclease free water	100 ml

Table 7.3. **Components of 2 M Acetic acid**

Components	Amount
Glacial acetic acid (60.05 g/L)	1.15 ml
De-ionised nuclease free water	9.85 ml

Table 7.4. **Substances used to prepare 10 mM Tris-HCl buffer, pH 8.0**

Components	Amount
Tris-acetate (121.14 g/L)	0.121 g
De-ionised nuclease free water	Up to 100 ml
Hydrochloric acid (HCl)	Adjusted to obtain pH of 8.0

7.3. Routinely used protocols

7.3.1. Preparation of template- positive Ion PGM Sphere Particles

The Ion OneTouch ™ Recovery Tube and Ion OneTouch ™ 2 Amplification Plate (connected to the disposable tube and injector) are put in their designated places with the disposable tubing inserted along the instrument's notches shaped. The injector is implanted straight in to the Ion OneTouch™ DL Injector Hub which is connected to the underlying recovery rotor.

Ion OneTouch™ Oil and OneTouch™ Recovery Solution are added (half full) to the two provided reagent tubes. Then new Sipper Tubes are attached to the corresponding ports and the filled reagent tubes are screwed firmly into their corresponding places. After ensuring

that the Waste Container is empty, both room temperatures equilibrated Ion PGM™ Template OT2 200 Reagent Mix and Ion PGM™ Template OT2 200 PCR Reagent B is vortexed and spun briefly to proceed with amplification step.

In the amplification step using a 1.5-ml Eppendorf LoBind tube, 25 µl of Nuclease Free Water, 500 µl of Ion PGM™ Template OT2 200 Reagent Mix, 300 µl of Ion PGM™ Template OT2 200 PCR Reagent B, 50 µl of briefly centrifuged and ice chilled PGM™ Template OT2 200 Enzyme Mix and 25 µl of the diluted library are mixed by briefly vortexed and centrifuged to create the amplification solution. This is followed by adding 100 µl of Ion PGM™ Template OT2 200 Ion Sphere™ Particles which is equilibrated to room temperature, vortexed, spun and mixed by pipetting up and down before its use. The total of 1000 µl amplification solution is vortexed for 5 sec before pipetting it to the sample port of the Ion PGM™ One Touch Plus Reaction Filter Assembly. Then 1500 µl of Reaction Oil is added through the same port. With the sample port of the assembly oriented to the right hand direction, the above Reaction Filter Assembly is carefully inverted to face down the three ports on the surface of the assembly. Then the three ports are inserted to the specified holes on the Ion OneTouch Instrument and the assembly is firmly seated. On the home screen the option Run is pressed. In the drop-down menu, the used template kit is selected among the list of predefined template kit lists and the run is started.

After completion of the amplification reaction on the Ion Sphere Particles, using the on screen prompts, final spin is performed. The ISP is collected from each recovery tube by removal of 50 µl of the supernatant from the surface and the opposite side of the ISP pellet using a pipette. This is followed by dispensing the ISP in the remaining solution by pipetting up and down. The product of up to 100 µl is transferred in to a new 1.5 ml Epppendorf LoBind tube, 1 ml of Ion OneTouch™ Wash Solution is added afterwards. The ISPs either ware stored at 4 ^{o}C for 1-2 days or used immediately. In any occasion, the Eppendorf tube containing ISP in the Wash solution is centrifuged at 15,500 rpm for 2.5 minutes, all the supernatant is removed except for 100 µl at the bottom of the tube which is used to re-suspend the ISPs using a pipette.

Ion OneTouch™ ES instrument is used to isolate the template enriched ISP from non-templated ISPs. This instrument uses automated magnetic bead technology for this purpose.

Initially, fresh Melt-Off Solution is prepared by combining 280 µl Tween^{R} Solution and 40 µl of 1 M NaOH. Then, 13 µl of vortexed Dynabeads MyOne™ Streptavidin C1 magnetic beads is put in a 1.5-ml LoBind Eppendorf Tube which is transferred to a magnetic rack until a clear supernatant is formed; subsequently the supernatant is discarded using a pipette. The bead is resuspended in 130 µl of MyOne™ Beads Wash Solution by vortexing and brief centrifugation. An 8-well strip from the Ion OneTouch™ kit is obtained, and placed in right side of Ion OneTouch™ ES instrument slot with left side orientation of the square-shaped tab of the strip. Following this, the 100 µl template enriched ISP sample is added to Well 1, while the 130 µl resuspended MyOne™ Beads to Well 2. Each of Well 3, 4, and 5 is loaded with 300 µl of Ion OneTouch™ Wash Solution. Wells 6 and 8 are left empty and Well 7 loaded with 300 µl of fresh Melt-Off Solution.

The Ion OneTouch™ ES instrument is prepared by placing a new tip on the Tip Arm and a fresh 0.2-ml PCR tube with its cap open in the base of the Tip Loader before initialising the run.

Following completion of this step, the Template-Positive Ion Sphere™ Particles accumulated in the 0.2 ml-PCR tube is mixed with 5 µl of Control Ion Sphere Particle and 100 µl of Annealing Buffer and is spun at 15,500 × g for 2 minutes. The supernatant is carefully removed except 15 µl is left in the tube. After adding 12 µl of Sequencing Primer, the mixture is pipetted up and down multiple times to disrupt the ISP pellet. To anneal the primers, the tube is put in a thermal cycler programmed with a heated lid and 95 ^{o}C for 2 minutes followed by 37 ^{o}C for 2 minutes.

7.3.2. Creating a planned run for Ion PGM system

Information about application of the PGM runs can be pre-saved for sequencing under the same conditions. It includes type of template (DNA or RNA), name of kit, and number of flows, barcodes and reference files. It can be done following the two steps below.

1. Log into the Torrent Browser to access the server connected to the PGM system.
2. Select Plan tab, then Templates and select the application type and then press Plan Run which can be reviewed using the Review tab or edited using Plan New Run tab

7.3.3. Cleaning of Ion PGM system

7.3.3.1. Cleaning the PGM machine with 18 MΩ water

Cleaning the system is necessary when the system is in daily use after every 2 runs or if > 10 hours but less than 48 hours have elapsed between the last cleaning/ initialisation and the start of a run. Before cleaning set up, the old sipper tubes and an old chip should be left in place. The 2 empty cleaning bottles provided are rinsed with 100 ml of fresh 18 MΩ water twice.
The sipper tube in W1 is rinsed with the same type of water put in a squirt bottle.

The cleaning bottles are placed in W2 and W3 positions and a collection tray below the sipper tubes in the dNTP positions. The tab Clean with 18 MΩ water is pressed to begin the cleaning process. After completion of the process, all the bottles and sipper tubes from W1, W2 and W3 positions are removed while the collection tray and the reagent sipper tubes are left in place.

7.3.3.2. Cleaning the PGM machine with Chlorite

This cleaning is performed once per week, unless the instrument has not been used since the last cleaning with Chlorite and whenever the machine is left with reagents for more than 48 hr.

1. Evacuate any remaining solution from each cleaning bottle and rinse twice with fresh 18 MΩ water.

2. In a large clean glass bottle dissolve PGM (Chlorite) Cleaning Tablet in 1 L of 18 MΩ water. Then add 1 mL of freshly prepared 1 M NaOH (prepared by dilution with 18 MΩ water) and filter the solution using 0.22-0.45 µm filter which should be used within 2 hours, 250 ml of this solution is then added to a 250 ml cleaning bottle.
3. The outside of sipper tube in W1 position is rinsed with 18 MΩ water. The other cleaning bottles are placed in W2 and W3 positions with a collection tray underneath the sipper tubes in the dNTP positions.
4. Press the Clean with Chlorite on the touch screen and press next to start cleaning.
5. After completion of the process, remove all the bottles and sipper tubes from W1, W2 and W3 positions while leave the collection tray and the reagent sipper tubes in place.

7.3.4. Sequencing on the Ion Torrent PGM

Cleaning of the Ion Torrent PGM machine (Life Technologies, UK) is performed with either 18 MΩ water alone or with Chlorite followed by 18 MΩ water washing.

Before initialising the Ion PGM system, The Wash 2 Bottle is filled with 18 MΩ water up to the upper marked Mold line, followed by pouring the entire content of Ion PGM™ Sequencing 200 v2 W2 and adding 70 µl of freshly prepared 100 mM NaOH solution. The bottle is then capped and the mixture is mixed by inverting the bottle several times. 350 µl of freshly prepared 100 mM NaOH is added to the Wash 1 Bottle and 50 ml Ion PGM™ Sequencing 200 v2 1 x W3 Solution is poured into Wash 3 Bottle. Following the onscreen instructions, the machine is initialised. After thawing each dNTP stock solution on ice, vortexing and brief spinning is performed. Carefully with changing gloves for handling each dNTP container, 20 µl of each dNTP stock solution is pipetted into its respective Reagent Bottle. The old dNTP sipper tubes are carefully replaced by new sipper tubes and the dNTP reagent bottles are firmly tightened to their corresponding places to complete the initialisation.

Chip Checking is performed to ensure that the machine can recognise and read it by scanning the new chip barcode and placing the chip on to the chip ground plate with bare

hands to avoid damage by electrostatic discharge. When this step is completed, the old chip is inserted to completely evacuate the waste bottle according to the touch screen instructions. To bind Sequencing Polymerase to the template positive-ISPs annealed to sequencing primer, 3 µl of Ion PGM™ Sequencing 200 v2 Polymerase is added to the 27 µl previously prepared template enriched- ISP sample and mixed by pipetting up and down. This mixture is left at room temperature for 5 minutes.

Ion 318™ Chip v2 (Life Technologies, UK) that passed the chip checking step, is taken out and tilted to 45 degree angle to bring down the loading port to lower most position. Withdrawal of any liquid in the chip is performed by a pipette through the circular loading port. For complete removal of any residual liquid, the chip is centrifuged upside down for 5 seconds using an adapter bucket. With the chip placed in a dry clean adapter bucket, the entire sample (~30 µl of template enrich ISPs) is loaded slowly with a firmly inserted pipette tip through the loading port to avoid introducing air bubble. The chip is transferred to the MiniFuge and centrifugation is performed for 30 seconds with the chip tab pointing toward the centre of the MiniFuge. Then, with a pipette set to 30 µl, the sample inside the chip is pipetted in and out slowly while the chip is tilted to 45 degrees with the loading port in lower most position.

Following this, the chip is centrifuged for another 30 seconds with the chip tab pointing away from the centre of the MiniFuge. Any residual sample is slowly removed from the loading port by a pipette with the chip at a 45-degree angle. After completion of the chip loading, the sequencing by 500 flow cycle run is performed following the displayed onscreen prompts.

7.3.5. Summary of somatic non-synonymous variants identified in serial NGS study of 23 CLL cases

Table 7.5. **Detailed information of somatic non-synonymous variants identified in longitudinal study**

Cases	Variant information					VAF% at follow-up time points			
	Gene (sub-clone)	Chr. Location (Hg19)	Ref	Var	Aminoacid change	TP1	TP2	TP3	TP4
CLL-1	*LRP1B*	2:141259338	C	T	p.R2923K	0	6		
	SF3B1 (A)	2:198266834	T	C	p.K700E	1	6		
	SF3B1 (B)	2:198267360	T	A	p.K666M	4	14		
	PCLO	7:82581658	G	T	P.P2871T	0	6		
	TP53 (A)	17:7577079	C	A	p.E278X	3	10		
	TP53 (B)	17:7577118	C	G	p.V274L	2	4		
	TP53 (C)	17:7576891	T	A	p.K319X	7	25		
	TP53 (D)	17:7577580	T	C	p.Y234C	1	3		
CLL-2	*XPO1*	2:61719472	C	T	p.E571K	29.8	25.3		
	LRP1B	2:141457824	A	G	p.V2265A	3.9	0		
	PCLO	7:82581607	C	T	p.V2888I	49.3	48.6		
	TP53	17:7577551	C	T	p.G244S	14.9	38.9		
CLL-3	*SF3B1*	2:198267699	G	A	p.R594X	5	0	4	
	TP53	17:7577568	C	T	p.C238Y	1	1.63	42.6	
CLL-4	*SF3B1*	2:198267385	A	G	p.W685R	49.4	49.8		
	BIRC3	11:102207709	A	-	p.V565fs	1.3	12.7		
	TP53	17:7577100	T	C	p.R280G	74.3	74.1		
CLL-5	*SF3B1*	2:198267491	C	A	p.E622D	40.4	47.4		
	PCLO (A)	7:82784834-35	-	Ins30N	p.Q374_Q375Ins10	24.6	23.5		
	PCLO (B)	7:82784832	C	G	p.Q375H	36.5	36		
	TP53	17:7577538	C	T	P.R248Q	64.8	83.6		
CLL-6	*SF3B1*	2:198266834	T	C	p.K700E	44.6	52.8		
	TP53	17:7578394	T	C	p.H179R	4.4	95.6		
	TP53	17:7578457	C	A	p.R158L	3.5	1		
CLL-7	*FBXW7*	4:153247330	C	G	p.G491A	0	0	9.9	
	TP53 (A)	17:7578525	G	C	p.C135W	35.6	62	31.2	
	TP53 (B)	17:7578211	C	A	p.R213L	13.7	14	54.3	
CLL-8	*SF3B1*	2:198268383	G	A	p.R549C	3.5	8.4	2	
	PCLO	7:82508679	G	A	p.A4543V	14.6	8	3.6	
	TP53 (A)	17:7572969-76	8N	-	p.HPP380X	41.2	2.7	36.8	
	TP53 (B)	17:7576854	T	C	p.Q331R	48.9	44.8	44	
	TP53 (C)	17:7578413	C	A	p.V173L	2	2.1	11	
	TP53 (D)	17:7577121	G	C	p.R273G	1	3.7	0	

Table 7.5. (continued)

Cases	Variant information					VAF% at follow-up time points			
	Gene (sub-clone)	Chr. Location (Hg19)	Ref	Var	Aminoacid change	TP1	TP2	TP3	TP4
CLL-10	*LRP1B*	2:141032152	T	A	p.Y4328F	0	3.9		
	SF3B1	2:198269834	A	T	p.L502X	2	5.5		
	NOTCH1 (A)	9:139390945	G	A	p.Q2416X	3.5	20.7		
	NOTCH1 (B)	9:139390649-50	CT	-	p.F2482Ffs*2	4	11.3		
	ATM (A)	11:108106443	T	A	p.D126E	53.7	50.9		
	ATM (B)	11:108175530-31	-	T	p.H1876fs	7.5	4.5		
CLL-11	*ATM* (A)	11:108121763	G	A	p.W524X	37.2	40.2		
	ATM (B)	11:108213973	G	A	p.G2765S	38.2	37.5		
CLL-12	*LRP1B*	2:141359103	C	T	p.R2302Q	3.2	0	0	
	SF3B1	2:198266834	T	C	p.K700E	10	20.8	29.9	
	ATM	11:108186598	T	C	p.Y2019H	97.6	96.5	98.4	
CLL-13	*XPO1*	2:61719472	C	T	p.E571K	36	30	24	
	ATM	11:108236179	G	C	p.A3039P	11	0	10	
CLL-14	*PCLO* (A)	7:82784834-35	-	Ins30N	p.Q374_Q375Ins10	42	41.4		
	PCLO (B)	7:82784832	C	G	p.Q375H	38.6	37.2		
	ATM	11:108236179	G	C	p.A3039P	0	6.6		
	SAMHD1 (A)	20:35526885-86	-	A	p.K523X	16	24.4		
	SAMHD1 (B)	20:35580045	A	T	p.M1K	58	71.9		
CLL-15	*ATM (A)*	11:108143528	T	G	p.L1078R	14.7	18.1		
	ATM (B)	11:108168106-07	-	C	p.L1668fs	14.4	12.7		
	ATM (C)	11:108236179	G	C	p.A3039P	0	5.6		
	NOTCH1	9:139390649-50	CT	-	p.F2482Ffs*2	10	8.6		
CLL-17	*SF3B1*	2:198269834	A	T	p.L502X	8	6.6	2	2
	ATM (A)	11:108186599	A	G	p.Y2019C	41	46.8	50.8	48.1
	ATM (B)	11:108216582-86	5N	-	p.W2845fs	46	47.5	47.9	48.3
CLL-18	*ATM*	11:108198445-54	10N	-	p.E2351fs	31	50.5		
CLL-19	*FBXW7*	4:153249384	C	A	p.R465L	5.3	4.4		
	CHD2 (A)	15:93499738	A	T	p.H620L	21.3	27.6		
	CHD2 (B)	15:93499735	C	G	p.A619G	21.3	29.5		
	SAMHD1	20:35545207	T	-	p.N327fs	7	21.2		
CLL-20	*SF3B1* (A)	2:198267699	G	A	p.R594X	0	5.4		
	SF3B1 (B)	2:198267370	T	C	p.T663A	54.9	80.2		
CLL-21	*SF3B1*	2:198266611	C	T	p.G742D	22.1	28.6	40.6	
CLL-22	*LRP1B*	2:141032021	T	A	p.N4372Y	41.7	48.6	49	
	ATM	11:108098521	A	T	p.K31X	4.1	0	0	
CLL-23	*MYD88*	3:38182032	C	G	p.S219C	4	0		
	SAMHD1	20:35526319	C	T	p.R551Q	49.1	49.4		
CLL-25	*PCLO*	7:82784834-35	-	Ins30N	p.Q374_Q375Ins10	41.3	38.2		
CLL-26	*HIST1H1E*	6:26156965	C	T	p.A116V	39	47.7	47.5	
	PCLO	7:82390770	T	C	p.K5016R	44	38.3	59.2	

7.3.6. Supplementary data of CNA test using CytoSNP 850K array

Table 7.6. **A summary of CNA detected with the SNP array in the 14 CLL samples studied**

CLL Samples	Type	Start Cyto	End Cyto	Chr	Start	End	Size (bp)	Comments
CLL-1	LOSS	4q21.1	4q21.3	4	76,692,898	87,004,521	10,311,624	
	LOSS	13q14.2	13q14.3	13	48,766,751	51,825,548	3,058,798	
CLL-3	LOSS	13q14.2	13q14.3	13	50,637,548	51,496,077	858,530	Biallelic
	LOSS	17p13.3	17p11.2	17	12,344	21,810,779	21,798,436	
	LOSS	18q22.1	18q23	18	65,709,017	77,982,456	12,273,440	
CLL-4	LOSS	4p16.3	4p15.1	4	49,450	32,515,265	32,465,816	
	LOSS	4q34.2	4q35.2	4	177,100,644	190,915,650	13,815,006	
	GAIN	7p22.3	7q36.3	7	44,935	159,126,310	159,081,376	Trisomy 7
	GAIN	8q21.3	8q24.3	8	93,187,902	146,293,414	53,105,512	
	LOSS	11q22.1	11q24.1	11	100,729,841	122,526,501	21,796,660	
	LOSS	13q14.2	13q22.2	13	48,184,645	76,104,956	27,920,312	
	LOH	17p13.3	17p13.1	17	12,344	9,163,144	9,150,800	Mosaic LOH 17p
	LOH	18q11.2	18q23	18	22,177,459	77,893,683	55,716,225	Mosaic LOH 18q
CLL-7 Latest	LOSS	2p25.3	2p16.1	2	14,238	58,795,436	58,781,199	
	LOSS	2p13.2	2p11.2	2	73,361,948	83,633,730	10,271,783	
	LOSS	6p25.2	6p24.3	6	3,077,141	9,275,594	6,198,454	
	LOSS	6p22.3	6p22.3	6	16,998,804	19,017,212	2,018,409	
	LOSS	6p21.33	6p21.2	6	30,770,000	38,728,142	7,958,143	
	LOSS	6p21.1	6p11.2	6	45,345,899	58,630,693	13,284,795	
	LOSS	6q11.1	6q12	6	61,891,118	67,977,326	6,086,209	
	LOSS	6q13	6q15	6	74,166,098	88,152,535	13,986,438	
	LOSS	7q21.11	7q21.12	7	79,585,451	87,226,690	7,641,240	
	LOSS	7q31.1	7q36.3	7	112,071,795	159,126,310	47,054,516	
	LOSS	8p23.3	8p12	8	164,984	35,042,190	34,877,207	
	LOSS	10q24.1	10q26.3	10	99,267,639	135,477,883	36,210,245	
	GAIN	11q22.3	11q25	11	107,617,257	134,934,063	27,316,807	
	LOSS	17p13.3	17p11.2	17	12,344	20,117,151	20,104,807	
	GAIN	19p13.3	19p13.11	19	5,830,302	18,342,477	12,512,176	
	LOSS	19p13.11	19p11	19	18,347,539	24,487,350	6,139,812	
CLL-7 Earlier	LOSS	2p25.3	2p21	2	14,238	46,082,995	46,068,757	
	LOSS	2p16.3	2p16.1	2	48,771,605	59,409,979	10,638,375	
	LOSS	2p13.1	2p11.2	2	73,556,155	84,249,446	10,693,292	
	LOSS	6p25.2	6p24.3	6	3,205,325	9,745,037	6,539,713	
	LOSS	6p21.33	6p21.2	6	30,724,430	38,723,575	7,999,146	
	LOSS	6p21.1	6q12	6	44,568,740	66,916,257	22,347,518	
	LOSS	6q13	6q15	6	74,217,278	88,021,118	13,803,841	
	LOSS	7q21.11	7q21.12	7	80,512,536	86,608,786	6,096,251	
	LOSS	7q31.1	7q32.1	7	112,044,527	127,479,262	15,434,736	
	LOSS	7q32.3	7q36.3	7	131,304,940	159,126,310	27,821,370	
	LOSS	8p23.3	8p11.23	8	164,984	36,646,217	36,481,234	
	LOSS	17p13.3	17p11.2	17	12,344	21,246,375	21,234,032	

Table 7.6. (continued)

CLL Samples	Type	Start Cyto	End Cyto	Chr	Start	End	Size (bp)	Comments
CLL-10	LOH	9q21.11	9q34.3	9	68838522	141066491	72227969	Mosaic LOH 9q
	GAIN	12p13.33	12q24.33	12	191,619	133,777,645	133,586,027	Trisomy 12
CLL-11	LOSS	8p23.3	8p12	8	164,984	32,141,515	31,976,532	
	GAIN	8q21.11	8q24.3	8	77,059,481	146,293,414	69,233,934	
	LOSS	13q14.11	13q14.11	13	40,825,707	41,797,397	971,691	
	LOSS	13q14.2	13q14.3	13	50,390,096	51,825,548	1,435,453	Biallelic
CLL-12	LOSS	11q21	11q23.3	11	94,388,085	115,638,755	21,250,671	
	LOSS	13q14.2	13q14.3	13	50,214,959	51,462,799	1,247,841	
CLL-17	GAIN	6p25.3	6p25.1	6	165,632	4,349,814	4,184,183	
	LOSS	6q14.1	6q27	6	83,704,533	170,919,470	87,214,938	
	LOSS	13q14.2	13q14.3	13	48,684,856	51,723,601	3,038,746	Deleted 13q14
	LOSS	18p11.32	18p11.21	18	13,034	14,753,891	14,740,858	
CLL-19	GAIN	12p13.33	12q24.33	12	191,619	133,777,645	133,586,026	
CLL-20	LOH	2q14.3	2q37.3	2	129,819,469	243,048,760	113,229,291	Mosaic LOH 2q
	LOSS	6q14.1	6q21	6	79,911,290	106,815,036	26,903,747	
	GAIN	12p13.33	12q24.33	12	191,619	133,777,645	133,586,027	Trisomy 12
	LOH	20q11.21	20q13.33	20	30686423	62912463	32226040	20q Complete LOH
CLL-21	NA	NA	NA	NA	Na	NA	NA	No CNA >5Mb detected
CLL-23	LOSS	12p13.2	12p13.1	12	10,925,439	14,337,924	3,412,485	
CLL-26	LOSS	7q33	7q34	7	134,746,551	140,012,638	5,266,088	
	LOH	11q13.2	11q25	11	65,962,002	134,934,063	68,972,061	Mosaic LOH 11q
	GAIN	12p13.33	12q24.33	12	191,619	133,777,645	133,586,027	Trisomy 12

References

1. *Cancer Research Uk, Chronic lymphocytic leukaemia (CLL) statistics. 2011, Cancer Research UK: UK.*

2. *Morrison, VA and Nowakowski, GS, Chronic lymphocytic leukemia. American Society of Hematology Self-Assessment Program, 2013: (2013)p. 579-95.*

3. *Demir, HA, Bayhan, T, Uner, A, Kurtulan, O, Karakus, E, Emir, S, et al., Chronic lymphocytic leukemia in a child: a challenging diagnosis in pediatric oncology practice. Pediatric Blood Cancer, 2014. 61(5): p. 933-35.*

4. *Eichhorst, B, Dreyling, M, Robak, T, Montserrat, E, Hallek, M, and Group, EGW, Chronic lymphocytic leukemia: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annual Oncology, 2011. 22(6): p. vi50-4.*

5. *Goldin, LR, Björkholm, M, Kristinsson, SY, Turesson, I, and Landgren, O, Elevated risk of chronic lymphocytic leukemia and other indolent non-Hodgkin's lymphomas among relatives of patients with chronic lymphocytic leukemia. Haematologica, 2009. 94(5): p. 647-53.*

6. *Lebien, TW and Tedder, TF, B lymphocytes: how they develop and function. Blood, 2008. 112(5): p. 1570-80.*

7. *Tonegawa, S, Somatic Generation of Antibody Diversity. Nature, 1983. 302: p. 575-81.*

8. *Osmond, DG, Population Dynamics of Bone Marrow B- Lymphocytes. Immunological Reviews, 1986. 93: p. 103-24.*

9. *Pancer, Z and Cooper, MD, The evolution of adaptive immunity. Annual Review Immunology, 2006. 24: p. 497-518.*

10. Blanchard-Rohner, G, Pulickal, AS, Jol-Van Der Zijde, CM, Snape, MD, and Pollard, AJ, Appearance of peripheral blood plasma cells and memory B cells in a primary and secondary immune response in humans. Blood, 2009. 114 (24): p. 4998-5002.

11. Ziqiang Li, CJw, et.al, The generation of antibody diversity through somatic hypermutation and class switch recombination. Gene and Developement, 2004. 18: p. 1-11.

12. Peakman, M, Basic and Clinical Immunology. 1997, USA: Churchill Livingstone. p. 78-84

13. Rose-Zerilli, MJ, Forster, J, Parry, M, Parker, A, Gardiner, A, Davies, Z, et al., Mutations In XPO1, POT1, BIRC3 and FBXW7 collectively Predict Poor Outcome At Diagnosis In CLL and MBL Independent From The SF3B1 and NOTCH1 Status. Blood, 2013. 122(21): p. 4138-38.

14. Chiorazzi, N and Ferrarini, M, Cellular origin(s) of chronic lymphocytic leukemia: cautionary notes and additional considerations and possibilities. Blood, 2011. 117(6): p. 1781-91.

15. Zhang, S and Kipps, TJ, The Pathogenesis of Chronic Lymphocytic Leukemia. Annual Review of Pathology: Mechanisms of Disease, 2014. 9(1): p. 103-18.

16. Fais, F, Ghiotto, F, Hashimoto, S, Sellars, B, Valetto, A, Allen, SL, et al., Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. Journal of Clinical Investigation, 1998. 102(8): p. 1515-25.

17. Zupo, S, Isnardi, L, Megna, M, Massara, R, Malavasi, F, Dono, M, et al., CD38 expression distinguishes two groups of B-cell chronic lymphocytic leukemias with different responses to anti-IgM antibodies and propensity to apoptosis. Blood, 1996. 88(4): p. 1365-74.

18. *Hamblin, TJ, Davis, Z, Gardiner, A, Oscier, DG, and Stevenson, FK, Unmutated Ig VH Genes Are Associated With a More Aggressive Form of Chronic Lymphocytic Leukemia. Blood, 1999. 94(6): p. 1848-54.*

19. *Gounari, M, Ntoufa, S, Apollonio, B, Papakonstantinou, N, Ponzoni, M, Chu, CC, et al., Excessive antigen reactivity may underlie the clinical aggressiveness of chronic lymphocytic leukemia stereotyped subset #8. Blood, 2015. 125(23): p. 3580-87.*

20. *Wiestner, A, Rosenwald, A, Barry, TS, Wright, G, Davis, RE, Henrickson, SE, et al., ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. Blood, 2003. 101(12): p. 4944-51.*

21. *Hallek, M, Cheson, BD, Catovsky, D, Caligaris-Cappio, F, Dighiero, G, Dohner, H, et al., Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. Blood, 2008. 111(12): p. 5446-56.*

22. *Rozman, C and Montserrat, E, Chronic Lymphocytic Leukemia. New England Journal of Medicine, 1995. 333(16): p. 1052-57.*

23. *Van Bockstaele, F, Verhasselt, B, and Philippé, J, Prognostic markers in chronic lymphocytic leukemia: A comprehensive review. Blood Reviews, 2009. 23(1): p. 25-47.*

24. *WHO Classification of Tumours of Haematopoeitic and Lymphoid Tissues. World Health Organization Classification of Tumours, ed. C.E. Swerdlow S., Harris N., Jaffe E., Pileri S.,Stein H, Lyon: International Agency for Research on Cancer 2008,4th edition, Vol.2.*

25. *Sheetal Desai, P, and Javier Pinilla-Ibarz, MD, PhD, Front-Line Therapy for Chronic Lymphocytic Leukemia. Cancer Control, 2012. 19(1): p. 26-36.*

26. Panasci, L, Paiement, J-P, Christodoulopoulos, G, Belenkov, A, Malapetsa, A, and Aloyz, R, Chlorambucil Drug Resistance in Chronic Lymphocytic Leukemia: The Emerging Role of DNA Repair. Clinical Cancer Research, 2001. 7(3): p. 454-61.

27. Catovsky, D, Richards, S, Matutes, E, Oscier, D, Dyer, MJS, Bezares, RF, et al., Assessment of fludarabine plus cyclophosphamide for patients with chronic lymphocytic leukaemia (the LRF CLL4 Trial): a randomised controlled trial. The Lancet, 2007. 370(9583): p. 230-39.

28. Goede , V, Fischer , K, Busch , R, Engelke , A, Eichhorst , B, Wendtner , CM, et al., Obinutuzumab plus Chlorambucil in Patients with CLL and Coexisting Conditions. New England Journal of Medicine, 2014. 370(12): p. 1101-10.

29. Woyach, JA and Johnson, AJ, Targeted therapies in CLL: mechanisms of resistance and strategies for management. Blood, 2015. 126(4): p. 471-77.

30. Johnson, GG, Sherrington, PD, Carter, A, Lin, K, Liloglou, T, Field, JK, et al., A Novel Type of p53 Pathway Dysfunction in Chronic Lymphocytic Leukemia Resulting from Two Interacting Single Nucleotide Polymorphisms within the p21 Gene. Cancer Research, 2009. 69(12): p. 5210-17.

31. Flinn, IW, Neuberg, DS, Grever, MR, Dewald, GW, Bennett, JM, Paietta, EM, et al., Phase III Trial of Fludarabine Plus Cyclophosphamide Compared With Fludarabine for Patients With Previously Untreated Chronic Lymphocytic Leukemia: US Intergroup Trial E2997. Journal of Clinical Oncology, 2007. 25(7): p. 793-98.

32. Pettitt, AR, Mechanism of action of purine analogues in chronic lymphocytic leukaemia. British Journal of Haematology, 2003. 121(5): p. 692-702.

33. Pettitt, AR, Sherrington, PD, Stewart, G, Cawley, JC, Taylor, AMR, and Stankovic, T, p53 dysfunction in B-cell chronic lymphocytic leukemia: inactivation of ATM as an alternative toTP53 mutation. Blood, 2001. 98(3): p. 814-22.

34. Van Oers, MHJ, Kuliczkowski, K, Smolej, L, Petrini, M, Offner, F, Grosicki, S, et al., Ofatumumab maintenance versus observation in relapsed chronic lymphocytic leukaemia (PROLONG): an open-label, multicentre, randomised phase 3 study. The Lancet Oncology, 2015. 16(13): p. 1370-79.

35. Owen, CJ and Stewart, DA, Obinutuzumab for the treatment of patients with previously untreated chronic lymphocytic leukemia: overview and perspective. Therapeutic Advances in Hematology, 2015. 6(4): p. 161-70.

36. Laurenti, L, Innocenti, I, Autore, F, Sica, S, and Efremov, DG, New developments in the management of chronic lymphocytic leukemia: role of ofatumumab. OncoTargets and Therapy, 2016. 9: p. 421-29.

37. Pettitt, AR, Jackson, R, Carruthers, S, Dodd, J, Dodd, S, Oates, M, et al., Alemtuzumab in Combination With Methylprednisolone Is a Highly Effective Induction Regimen for Patients With Chronic Lymphocytic Leukemia and Deletion of TP53: Final Results of the National Cancer Research Institute CLL206 Trial. Journal of Clinical Oncology, 2012. 30(14): p. 1647-55.

38. Pettitt, AR, Polydoros, F, Dodd, J, Oates, M , Lin, K, Kalakonda, N, et al.,. Final results of the NCRI CLL210 trial of alemtuzumab, dexamethazone and lenaledomide in patients with high-risk CLL (original protocol). European Haematology Association 2016. Denimark: EHA Learning Centre.

39. Herman, SEM, Gordon, AL, Wagner, AJ, Heerema, NA, Zhao, W, Flynn, JM, et al., Phosphatidylinositol 3-kinase-δ inhibitor CAL-101 shows promising preclinical activity in chronic lymphocytic leukemia by antagonizing intrinsic and extrinsic cellular survival signals. Blood, 2010. 116(12): p. 2078-88.

40. Furman , RR, Sharman , JP, Coutre , SE, Cheson , BD, Pagel , JM, Hillmen , P, et al., Idelalisib and Rituximab in Relapsed Chronic Lymphocytic Leukemia. New England Journal of Medicine, 2014. 370(11): p. 997-1007.

41. Byrd , JC, Brown , JR, O'brien , S, Barrientos , JC, Kay , NE, Reddy , NM, et al., Ibrutinib versus Ofatumumab in Previously Treated Chronic Lymphoid Leukemia. New England Journal of Medicine, 2014. 371(3): p. 213-23.

42. Woyach, JA, Furman, RR, Liu, TM, Ozer, HG, Zapatka, M, Ruppert, AS, et al., Resistance Mechanisms for the Bruton's Tyrosine Kinase Inhibitor Ibrutinib. New England Journal of Medicine, 2014. 370(24): p. 2286-94.

43. Maddocks, KJ, Ruppert, AS, Lozanski, G, and Et Al., ETiology of ibrutinib therapy discontinuation and outcomes in patients with chronic lymphocytic leukemia. Journal of American Medical Assocaition Oncology, 2015. 1(1): p. 80-87.

44. Kitada, S, Andersen, J, Akar, S, Zapata, JM, Takayama, S, Krajewski, S, et al., Expression of apoptosis-regulating proteins in chronic lymphocytic leukemia: correlations with in vitro and in vivo chemoresponses. Blood, 1998. 91(9): p. 3379-89.

45. Davids, MS, Pagel, JM, Kahl, BS, Wierda, WG, Miller, TP, Gerecitano, JF, et al., Bcl-2 Inhibitor ABT-199 (GDC-0199) Monotherapy Shows Anti-Tumor Activity Including Complete Remissions In High-Risk Relapsed/Refractory (R/R) Chronic Lymphocytic Leukemia (CLL) and Small Lymphocytic Lymphoma (SLL). Blood, 2013. 122(21): p. 872-72.

46. Lapalombella, R, Sun, Q, Williams, K, Tangeman, L, Jha, S, Zhong, Y, et al., Selective inhibitors of nuclear export show that CRM1/XPO1 is a target in chronic lymphocytic leukemia. Blood, 2012. 120(23): p. 4621-34.

47. Chiorazzi, N, Implications of new prognostic markers in chronic lymphocytic leukemia. ASH Education Program Book, 2012. 2012(1): p. 76-87.

48. Källander, CFR, Simonsson, B, Hagberg, H, and Gronowitz, JS, Serum deoxythymidine kinase gives prognostic information in chronic lymphocytic leukemia. Cancer, 1984. 54(11): p. 2450-55.

49. Di Giovanni, S, Valentini, G, Carducci, P, and Giallonardo, P, Beta-2-Microglobulin Is a Reliable Tumor Marker in Chronic Lymphocytic Leukemia. Acta Haematologica, 1989. 81(4): p. 181-85.

50. Sarfati, M, Chevret, S, Chastang, C, Biron, G, Stryckmans, P, Delespesse, G, et al., Prognostic importance of serum soluble CD23 level in chronic lymphocytic leukemia. Blood, 1996. 88(11): p. 4259-64.

51. Christiansen, I, Sundström, C, and Tötterman, TH, Elevated serum levels of soluble vascular cell adhesion molecule-1 (sVCAM-1) closely reflect tumour burden in chronic B-lymphocytic leukaemia. British Journal of Haematology, 1998. 103(4): p. 1129-37.

52. Damle, RN, Wasil, T, Fais, F, Ghiotto, F, Valetto, A, Allen, SL, et al., Ig V Gene Mutation Status and CD38 Expression As Novel Prognostic Indicators in Chronic Lymphocytic Leukemia. Presented in part at the 40th Annual Meeting of The American Society of Hematology, held in Miami Beach, FL, December 4-8, 1998., 1999. 94(6): p. 1840-47.

53. Moreno, C and Montserrat, E, New prognostic markers in chronic lymphocytic leukemia. Blood Reviews, 2008. 22(4): p. 211-19.

54. Renata, W and David, O, Prognostic markers in chronic lymphocytic leukemia, in Advances in the Treatment of B-Cell Chronic Lymphocytic Leukemia. 2012, Future Medicine Ltd. p. 76-86.

55. Matsuda, F, Shin, EK, Nagaoka, H, Matsumura, R, Haino, M, Fukita, Y, et al., STRUCTURE AND PHYSICAL MAP OF 64 VARIABLE SEGMENTS IN THE 3(I)0.8-MEGABASE REGION OF THE HUMAN-IMMUNOGLOBULIN HEAVY-CHAIN LOCUS. Nature Genetics, 1993. 3(1): p. 88-94.

56. Murray, F, Darzentas, N, Hadzidimitriou, A, Tobin, G, Boudjogra, M, Scielzo, C, et al., Stereotyped patterns of somatic hypermutation in subsets of patients with chronic lymphocytic leukemia: implications for the role of antigen selection in leukemogenesis. Blood, 2008. 111(3): p. 1524-33.

57. Klein, U, Rajewsky, K, and Küppers, R, Human Immunoglobulin (Ig)M+IgD+ Peripheral Blood B Cells Expressing the CD27 Cell Surface Antigen Carry Somatically Mutated Variable Region Genes: CD27 as a General Marker for Somatically Mutated (Memory) B Cells. The Journal of Experimental Medicine, 1998. 188(9): p. 1679-89.

58. Stamatopoulos, K, Belessi, C, Hadzidimitriou, A, Smilevska, T, Kalagiakou, E, Hatzi, K, et al., Immunoglobulin light chain repertoire in chronic lymphocytic leukemia. Blood, 2005. 106(10): p. 3575-83.

59. Oscier, DG, Thompsett, A, Zhu, D, and Stevenson, FK, Differential Rates of Somatic Hypermutation in VH Genes Among Subsets of Chronic Lymphocytic Leukemia Defined by Chromosomal Abnormalities. Blood, 1997. 89(11): p. 4153-60.

60. Wiestner, A, Rosenwald, A, Barry, TS, Wright, G, Davis, RE, Henrickson, SE, et al., ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. Blood, 2003. 101(12): p. 4944-51.

61. Kienle, D, Benner, A, Kröber, A, Winkler, D, Mertens, D, Bühler, A, et al., Distinct gene expression patterns in chronic lymphocytic leukemia defined by usage of specific VH genes. Blood, 2006. 107(5): p. 2090-93.

62. Tschumper, RC, Geyer, SM, Campbell, ME, Kay, NE, Shanafelt, TD, Zent, CS, et al., Immunoglobulin diversity gene usage predicts unfavorable outcome in a subset of chronic lymphocytic leukemia patients. The Journal of Clinical Investigation, 2008. 118(1): p. 306-15.

63. Lin, K, Manocha, S, Harris, RJ, Matrai, Z, Sherrington, PD, and Pettitt, AR, High frequency of p53 dysfunction and low level of VH mutation in chronic lymphocytic leukemia patients using the VH3-21 gene segment. Blood, 2003. 102(3): p. 1145-46.

64. Thorsélius, M, Kröber, A, Murray, F, Thunberg, U, Tobin, G, Bühler, A, et al., Strikingly homologous immunoglobulin gene rearrangements and poor outcome in VH3-21-

using chronic lymphocytic leukemia patients independent of geographic origin and mutational status. Blood, 2006. 107(7): p. 2889-94.

65. *Tobin, G, Thunberg, U, Johnson, A, Thörn, I, Söderberg, O, Hultdin, M, et al., Somatically mutated Ig VH3-21 genes characterize a new subset of chronic lymphocytic leukemia. Blood, 2002. 99(6): p. 2262-64.*

66. *Hamblin, TJ, Orchard, JA, Ibbotson, RE, Davis, Z, Thomas, PW, Stevenson, FK, et al., CD38 expression and immunoglobulin variable region mutations are independent prognostic variables in chronic lymphocytic leukemia, but CD38 expression may vary during the course of the disease. Blood, 2002. 99(3): p. 1023-29.*

67. *Dürig, J, Naschar, M, Schmücker, U, Renzing-Köhler, K, Hölter, T, Hüttmann, A, et al., CD38 expression is an important prognostic marker in chronic lymphocytic leukaemia. Leukemia, 2002. 16(1): p. 30-35.*

68. *Ibrahim, S, Keating, M, Do, K-A, O'brien, S, Huh, YO, Jilani, I, et al., CD38 expression as an important prognostic factor in B-cell chronic lymphocytic leukemia. Blood, 2001. 98(1): p. 181-86.*

69. *Del Poeta, G, Maurillo, L, Venditti, A, Buccisano, F, Epiceno, AM, Capelli, G, et al., Clinical significance of CD38 expression in chronic lymphocytic leukemia. Blood, 2001. 98(9): p. 2633-39.*

70. *Damle, RN, Wasil, T, Fais, F, Ghiotto, F, Valetto, A, Allen, SL, et al., Ig V Gene Mutation Status and CD38 Expression As Novel Prognostic Indicators in Chronic Lymphocytic Leukemia Presented in part at the 40th Annual Meeting of The American Society of Hematology, held in Miami Beach, FL, December 4-8, 1998. Blood, 1999. 94(6): p. 1840-47.*

71. *Van Bockstaele, F, Verhasselt, B, and Philippé, J, Prognostic markers in chronic lymphocytic leukemia: A comprehensive review. Blood Reviews. 23(1): p. 25-47.*

72. Kane, LP, Lin, J, and Weiss, A, Signal transduction by the TCR for antigen. Current Opinion in Immunology, 2000. 12(3): p. 242-49.

73. Chen, L, Widhopf, G, Huynh, L, Rassenti, L, Rai, KR, Weiss, A, et al., Expression of ZAP-70 is associated with increased B-cell receptor signaling in chronic lymphocytic leukemia. Blood, 2002. 100(13): p. 4609-14.

74. Rassenti, LZ, Huynh, L, Toy, TL, Chen, L, Keating, MJ, Gribben, JG, et al., ZAP-70 Compared with Immunoglobulin Heavy-Chain Gene Mutation Status as a Predictor of Disease Progression in Chronic Lymphocytic Leukemia. New England Journal of Medicine, 2004. 351(9): p. 893-901.

75. Döhner, H, Stilgenbauer, S, Benner, A, Leupolt, E, Kröber, A, Bullinger, L, et al., Genomic aberrations and survival in chronic lymphocytic leukemia. New England Journal of Medicine, 2000. 343(26): p. 1910-16.

76. Pfeifer, D, Pantic, M, Skatulla, I, Rawluk, J, Kreutz, C, Martens, UM, et al., Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays. Blood, 2007. 109(3): p. 1202-10.

77. Gunnarsson, R, Staaf, J, Jansson, M, Ottesen, AM, Göransson, H, Liljedahl, U, et al., Screening for copy-number alterations and loss of heterozygosity in chronic lymphocytic leukemia—A comparative study of four differently designed, high resolution microarray platforms. Genes, Chromosomes and Cancer, 2008. 47(8): p. 697-711.

78. Schwaenen, C, Nessling, M, Wessendorf, S, Salvi, T, Wrobel, G, Radlwimmer, B, et al., Automated array-based genomic profiling in chronic lymphocytic leukemia: Development of a clinical tool and discovery of recurrent genomic alterations. Proceedings of the National Academy of Sciences of the United States of America, 2004. 101(4): p. 1039-44.

79. Chapiro, E, Leporrier, N, Radford-Weiss, I, Bastard, C, Mossafa, H, Leroux, D, et al., Gain of the short arm of chromosome 2 (2p) is a frequent recurring chromosome aberration in untreated chronic lymphocytic leukemia (CLL) at advanced stages. Leukemia Research, 2010. 34(1): p. 63-68.

80. Ouillette, P, Erba, H, Kujawski, L, Kaminski, M, Shedden, K, and Malek, SN, Integrated Genomic Profiling of Chronic Lymphocytic Leukemia Identifies Subtypes of Deletion 13q14. Cancer Research, 2008. 68(4): p. 1012-21.

81. Van Dyke, DL, Shanafelt, TD, Call, TG, Zent, CS, Smoley, SA, Rabe, KG, et al., A comprehensive evaluation of the prognostic significance of 13q deletions in patients with B-chronic lymphocytic leukaemia. British Journal of Haematology, 2010. 148(4): p. 544-50.

82. Dewald, GW, Brockman, SR, Paternoster, SF, Bone, ND, O'fallon, JR, Allmer, C, et al., Chromosome anomalies detected by interphase fluorescence in situ hybridization: correlation with significant biological features of B-cell chronic lymphocytic leukaemia. British journal of haematology, 2003. 121(2): p. 287-95.

83. Garg, R, Wierda, W, Ferrajoli, A, Abruzzo, L, Pierce, S, Lerner, S, et al., The prognostic difference of monoallelic versus biallelic deletion of 13q in chronic lymphocytic leukemia. Cancer, 2012. 118(14): p. 3531-37.

84. Liu, Y, Corcoran, M, Rasool, O, Ivanova, G, Ibbotson, R, Grander, D, et al., Cloning of two candidate tumor suppressor genes within a 10 kb region on chromosome 13q14, frequently deleted in chronic lymphocytic leukemia. Oncogene, 1997. 15(20): p. 2463-73.

85. Migliazza, A, Bosch, F, Komatsu, H, Cayanis, E, Martinotti, S, Toniato, E, et al., Nucleotide sequence, transcription map, and mutation analysis of the 13q14 chromosomal region deleted in B-cell chronic lymphocytic leukemia. Blood, 2001. 97(7): p. 2098-104.

86. Thai, T-H, Dysregulation of MicroRNA Expression and Human Diseases?, in From Nucleic Acids Sequences to Molecular Medicine, V.A. Erdmann and J. Barciszewski, Editors. 2012, Springer Berlin Heidelberg. p. 553-71.

87. Fabbri, M, Bottoni, A, Shimizu, M, Spizzo, R, Nicoloso, MS, Rossi, S, et al., Association of a microRNA/TP53 feedback circuitry with pathogenesis and outcome of B-cell chronic lymphocytic leukemia. Journal of American Medical Association, 2011. 305(1): p. 59-67.

88. Lin, K, Farahani, M, Yang, Y, Johnson, GG, Oates, M, Atherton, M, et al., Loss of MIR15A and MIR16-1 at 13q14 is associated with increased TP53 mRNA, de-repression of BCL2 and adverse outcome in chronic lymphocytic leukaemia. British Journal of Haematology, 2014. 167(3): p. 346-55.

89. Dal Bo, M, Rossi, FM, Rossi, D, Deambrogi, C, Bertoni, F, Del Giudice, I, et al., 13q14 Deletion Size and Number of Deleted Cells Both Influence Prognosis in Chronic Lymphocytic Leukemia. Genes Chromosomes & Cancer, 2011. 50(8): p. 633-43.

90. Fangazio, M, De Paoli, L, Rossi, D, and Gaidano, G, Predictive markers and driving factors behind Richter syndrome development. Expert Review Anticancer Therepy, 2011. 11(3): p. 433-42.

91. Parker, H, Rose-Zerilli, M, Parker, A, Chaplin, T, Wade, R, Gardiner, A, et al., 13q deletion anatomy and disease progression in patients with chronic lymphocytic leukemia. Leukemia, 2011. 25(3): p. 489-97.

92. Jeromin, S, Weissmann, S, Haferlach, C, Dicker, F, Bayer, K, Grossmann, V, et al., SF3B1 mutations correlated to cytogenetics and mutations in NOTCH1, FBXW7, MYD88, XPO1 and TP53 in 1160 untreated CLL patients. Leukemia, 2014. 28(1): p. 108-17.

93. Landau, DA, Tausch, E, Taylor-Weiner, AN, Stewart, C, Reiter, JG, Bahlo, J, et al., Mutations driving CLL and their evolution in progression and relapse. Nature, 2015. 526(7574): p. 525-30.

94. Döhner, H, Stilgenbauer, S, James, MR, Benner, A, Weilguni, T, Bentz, M, et al., 11q Deletions Identify a New Subset of B-Cell Chronic Lymphocytic Leukemia Characterized by Extensive Nodal Involvement and Inferior Prognosis. Blood, 1997. 89(7): p. 2516-22.

95. Guarini, A, Marinelli, M, Tavolaro, S, Bellacchio, E, Magliozzi, M, Chiaretti, S, et al., ATM gene alterations in chronic lymphocytic leukemia patients induce a distinct gene expression profile and predict disease progression. Haematologica, 2012. 97(1): p. 47-55.

96. Gately, DP, Hittle, JC, Chan, GKT, and Yen, TJ, Characterization of ATM Expression, Localization, and Associated DNA-dependent Protein Kinase Activity. Molecular Biology of the Cell, 1998. 9(9): p. 2361-74.

97. Stilgenbauer, S, Liebisch, P, James, MR, Schröder, M, Schlegelberger, B, Fischer, K, et al., Molecular cytogenetic delineation of a novel critical genomic region in chromosome bands 11q22.3-923.1 in lymphoproliferative disorders. Proceedings of the National Academy of Sciences, 1996. 93(21): p. 11837-41.

98. Joshi, AD, Dickinson, JD, Hegde, GV, Sanger, WG, Armitage, JO, Bierman, PJ, et al., Bulky lymphadenopathy with poor clinical outcome is associated with ATM downregulation in B-cell chronic lymphocytic leukemia patients irrespective of 11q23 deletion. Cancer Genetics and Cytogenetics, 2007. 172(2): p. 120-26.

99. Austen, B, Skowronska, A, Baker, C, Powell, JE, Gardiner, A, Oscier, D, et al., Mutation Status of the Residual ATM Allele Is an Important Determinant of the Cellular Response to Chemotherapy and Survival in Patients With Chronic Lymphocytic Leukemia Containing an 11q Deletion. Journal of Clinical Oncology, 2007. 25(34): p. 5448-57.

100. Fegan, C, Robinson, H, Thompson, P, Whittaker, JA, and White, D, Karyotypic evolution in CLL: identification of a new sub-group of patients with deletions of 11q and advanced or progressive disease. Leukemia, 1995. 9(12): p. 2003-08.

101. Grossmann, V, Kohlmann, A, Schnittger, S, Weissmann, S, Jeromin, S, Kienast, J, et al. Recurrent ATM and BIRC3 mutations in patients with chronic lymphocytic leukemia (CLL) and deletion 11q22-q23. in Blood. 2012. AMER SOC HEMATOLOGY 2021 L ST NW, SUITE 900, WASHINGTON, DC 20036 USA.

102. Rossi, D, Rasi, S, Spina, V, Bruscaggin, A, Monti, S, Ciardullo, C, et al., Integrated mutational and cytogenetic analysis identifies new prognostic subgroups in chronic lymphocytic leukemia. Blood, 2013. 121(8): p. 1403-12.

103. Rose-Zerilli, MJJ, Forster, J, Parker, H, Parker, A, Rodriguez, AE, Chaplin, T, et al., ATM mutation rather than BIRC3 deletion and/or mutation predicts reduced survival in 11q-deleted chronic lymphocytic leukemia, data from the UK LRF CLL4 trial. Haematologica, 2014. 99(4): p. 736-42.

104. Dohner, H, Fischer, K, Bentz, M, Hansen, K, Benner, A, Cabot, G, et al., p53 gene deletion predicts for poor survival and non-response to therapy with purine analogs in chronic B-cell leukemias. Blood, 1995. 85(6): p. 1580-89.

105. Zenz, T, Gribben, JG, Hallek, M, Döhner, H, Keating, MJ, and Stilgenbauer, S, Risk categories and refractory CLL in the era of chemoimmunotherapy. Blood, 2012. 119(18): p. 4101-07.

106. Zenz, T, Häbe, S, Denzel, T, Mohr, J, Winkler, D, Bühler, A, et al., Detailed analysis of p53 pathway defects in fludarabine-refractory chronic lymphocytic leukemia (CLL): dissecting the contribution of 17p deletion, TP53 mutation, p53-p21 dysfunction, and miR34a in a prospective clinical trial. Blood, 2009. 114(13): p. 2589-97.

107. May, P and May, E, Twenty years of p53 research: structural and functional aspects of the p53 protein. Oncogene, 1999. 18(53): p. 7621-36.

108. Zenz, T, Vollmer, D, Trbusek, M, Smardova, J, Benner, A, Soussi, T, et al., TP53 mutation profile in chronic lymphocytic leukemia: evidence for a disease specific profile from a comprehensive analysis of 268 mutations. Leukemia, 2010. 24(12): p. 2072-9.

109. Dicker, F, Herholz, H, Schnittger, S, Nakao, A, Patten, N, Wu, L, et al., The detection of TP53 mutations in chronic lymphocytic leukemia independently predicts rapid disease progression and is highly correlated with a complex aberrant karyotype. Leukemia, 2009. 23(1): p. 117-24.

110. Oscier, DG, Gardiner, AC, Mould, SJ, Glide, S, Davis, ZA, Ibbotson, RE, et al., Multivariate analysis of prognostic factors in CLL: clinical stage, IGVH gene mutational status, and loss or mutation of the p53 gene are independent prognostic factors. Blood, 2002. 100(4): p. 1177-84.

111. Gribben, JG, How I treat CLL up front. Blood, 2010. 115(2): p. 187-97.

112. Rossi, D, Rasi, S, Spina, V, Bruscaggin, A, Monti, S, Ciardullo, C, et al., Integrated mutational and cytogenetic analysis identifies new prognostic subgroups in chronic lymphocytic leukemia. Blood, 2013. 121(8): p. 1403-12.

113. Balatti, V, Bottoni, A, Palamarchuk, A, Alder, H, Rassenti, LZ, Kipps, TJ, et al., NOTCH1 mutations in CLL associated with trisomy 12. Blood, 2012. 119(2): p. 329-31.

114. Lopez, C, Delgado, J, Costa, D, Conde, L, Ghita, G, Villamor, N, et al., Different distribution of NOTCH1 mutations in chronic lymphocytic leukemia with isolated trisomy 12 or associated with other chromosomal alterations. Genes Chromosomes Cancer, 2012. 51(9): p. 881-9.

115. Porpaczy, E, Bilban, M, Heinze, G, Gruber, M, Vanura, K, Schwarzinger, I, et al., Gene expression signature of chronic lymphocytic leukaemia with Trisomy 12. European Journal of Clinical Investigation, 2009. 39(7): p. 568-75.

116. Kienle, DL, Korz, C, Hosch, B, Benner, A, Mertens, D, Habermann, A, et al., Evidence for distinct pathomechanisms in genetic subgroups of chronic lymphocytic leukemia revealed by quantitative expression analysis of cell cycle, activation, and apoptosis-associated genes. Journal of Clinical Oncology, 2005. 23(16): p. 3780-92.

117. Landau, DA, Carter, SL, Stojanov, P, Mckenna, A, Stevenson, K, Lawrence, MS, et al., Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell, 2013. 152(4): p. 714-26.

118. Athanasiadou, A, Stamatopoulos, K, Tsompanakou, A, Gaitatzi, M, Kalogiannidis, P, Anagnostopoulos, A, et al., Clinical, immunophenotypic, and molecular profiling of trisomy 12 in chronic lymphocytic leukemia and comparison with other karyotypic subgroups defined by cytogenetic analysis. Cancer Genetics and Cytogenetics, 2006. 168(2): p. 109-19.

119. Edelmann, J, Holzmann, K, Miller, F, Winkler, D, Bühler, A, Zenz, T, et al., High-resolution genomic profiling of chronic lymphocytic leukemia reveals new recurrent genomic alterations. Blood, 2012. 120(24): p. 4783-94.

120. Baliakas, P, Iskas, M, Gardiner, A, Davis, Z, Plevova, K, Nguyen-Khac, F, et al., Chromosomal translocations and karyotype complexity in chronic lymphocytic leukemia: A systematic reappraisal of classic cytogenetic data. American Journal of Hematology, 2014. 89(3): p. 249-55.

121. Haferlach, C, Dicker, F, Schnittger, S, Kern, W, and Haferlach, T, Comprehensive genetic characterization of CLL: a study on 506 cases analysed with chromosome banding analysis, interphase FISH, IgVH status and immunophenotyping. Leukemia, 2007. 21(12): p. 2442-51.

122. Ouillette, P, Fossum, S, Parkin, B, Ding, L, Bockenstedt, P, Al-Zoubi, A, et al., Aggressive Chronic Lymphocytic Leukemia with Elevated Genomic Complexity Is Associated with Multiple Gene Defects in the Response to DNA Double-Strand Breaks. Clinical Cancer Research, 2010. 16(3): p. 835-47.

123. *Van Den Neste, E, Robin, V, Francart, J, Hagemeijer, A, Stul, M, Vandenberghe, P, et al., Chromosomal translocations independently predict treatment failure, treatment-free survival and overall survival in B-cell chronic lymphocytic leukemia patients treated with cladribine. Leukemia, 2007. 21(8): p. 1715-22.*

124. *Mayr, C, Speicher, MR, Kofler, DM, Buhmann, R, Strehl, J, Busch, R, et al., Chromosomal translocations are associated with poor prognosis in chronic lymphocytic leukemia. Blood, 2006. 107(2): p. 742-51.*

125. *Cavazzini, F, Hernandez, JA, Gozzetti, A, Russo Rossi, A, De Angeli, C, Tiseo, R, et al., Chromosome 14q32 translocations involving the immunoglobulin heavy chain locus in chronic lymphocytic leukaemia identify a disease subset with poor prognosis. British Journal of Haematology, 2008. 142(4): p. 529-37.*

126. *Hruba, M, Dvorak, P, Weberova, L, and Subrt, I, Independent coexistence of clones with 13q14 deletion at reciprocal translocation breakpoint and 13q14 interstitial deletion in chronic lymphocytic leukemia. Leukemia & lymphoma, 2012. 53(10): p. 2054-62.*

127. *Struski, S, Helias, C, Gervais, C, Audhuy, B, Zamfir, A, Herbrecht, R, et al., 13q deletions in B-cell lymphoproliferative disorders: frequent association with translocation. Cancer Genetics and Cytogenetics, 2007. 174(2): p. 151-60.*

128. *Puiggros, A, Blanco, G, and Espinet, B, Genetic Abnormalities in Chronic Lymphocytic Leukemia: Where We Are and Where We Go. BioMed Research International, 2014. 2014: p. 13.*

129. *Pei, J, Jhanwar, SC, and Testa, JR, Chromothripsis in a Case of TP53-Deficient Chronic Lymphocytic Leukemia. Leukemia Research Reports, 2012. 1(1): p. 4-6.*

130. *Zhu, J, Zhang, S, Jiang, J, and Chen, X, Definition of the p53 Functional Domains Necessary for Inducing Apoptosis. Journal of Biological Chemistry, 2000. 275(51): p. 39927-34.*

131. *Ahmed, AA, Etemadmoghadam, D, Temple, J, Lynch, AG, Riad, M, Sharma, R, et al., Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. The Journal of Pathology, 2010. 221(1): p. 49-56.*

132. *Rivlin, N, Brosh, R, Oren, M, and Rotter, V, Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis. Genes & Cancer, 2011. 2(4): p. 466-74.*

133. *Bullock, AN and Fersht, AR, Rescuing the function of mutant p53. Nature Review Cancer, 2001. 1(1): p. 68-76.*

134. *Zenz, T, Vollmer, D, Trbusek, M, Smardova, J, Benner, A, Soussi, T, et al., TP53 mutation profile in chronic lymphocytic leukemia: evidence for a disease specific profile from a comprehensive analysis of 268 mutations. Leukemia, 2010. 24(12): p. 2072-79.*

135. *Zenz, T, Eichhorst, B, Busch, R, Denzel, T, Habe, S, Winkler, D, et al., TP53 mutation and survival in chronic lymphocytic leukemia. Journal of Clinical Oncology, 2010. 28(29): p. 4473-9.*

136. *Jeromin, S, Kern, W, Schabath, R, Alpermann, T, Nadarajah, N, Meggendorfer, M, et al. Modulation of the Clonal Composition in Relapsed CLL: A Study Based on Targeted Deep-Sequencing of ATM, BIRC3, NOTCH1, POT1, SF3B1, SAMHD1 and TP53. in American Society of Haematology annual meeting. 2015. Orlando, Florida: American Socity of Hematology.*

137. *Malcikova, J, Stano-Kozubik, K, Tichy, B, Kantorova, B, Pavlova, S, Tom, N, et al., Detailed analysis of therapy-driven clonal evolution of TP53 mutations in chronic lymphocytic leukemia. Leukemia, 2015. 29(4): p. 877-85.*

138. *Rossi, D, Khiabanian, H, Spina, V, Ciardullo, C, Bruscaggin, A, Famà, R, et al., Clinical impact of small TP53 mutated subclones in chronic lymphocytic leukemia. Blood, 2014. 123(14): p. 2139-47.*

139. Trbusek, M, Smardova, J, Malcikova, J, Sebejova, L, Dobes, P, Svitakova, M, et al., Missense Mutations Located in Structural p53 DNA-Binding Motifs Are Associated With Extremely Poor Survival in Chronic Lymphocytic Leukemia. Journal of Clinical Oncology, 2011. 29(19): p. 2703-08.

140. Negrini, S, Gorgoulis, VG, and Halazonetis, TD, Genomic instability - an evolving hallmark of cancer. Nature Reviews Molecular Cell Biology, 2010. 11(3): p. 220-28.

141. Saito, Si, Goodarzi, AA, Higashimoto, Y, Noda, Y, Lees-Miller, SP, Appella, E, et al., ATM Mediates Phosphorylation at Multiple p53 Sites, Including Ser46, in Response to Ionizing Radiation. Journal of Biological Chemistry, 2002. 277(15): p. 12491-94.

142. Austen, B, Powell, JE, Alvi, A, Edwards, I, Hooper, L, Starczynski, J, et al., Mutations in the ATM gene lead to impaired overall and treatment-free survival that is independent of IGVH mutation status in patients with B-CLL. Blood, 2005. 106(9): p. 3175-82.

143. Gumy-Pause, F, Wacker, P, and Sappino, AP, ATM gene and lymphoid malignancies. Leukemia, 2003. 18(2): p. 238-42.

144. Stankovic, T, Weber, P, Stewart, G, Bedenham, T, Murray, J, Byrd, PJ, et al., Inactivation of ataxia telangiectasia mutated gene in B-cell chronic lymphocytic leukaemia. The Lancet, 1999. 353(9146): p. 26-29.

145. Navrkalova, V, Sebejova, L, Zemanova, J, Kminkova, J, Kubesova, B, Malcikova, J, et al., ATM mutations uniformly lead to ATM dysfunction in chronic lymphocytic leukemia: application of functional test using doxorubicin. Haematologica, 2013. 98(7): p. 1124-31.

146. Skowronska, A, Parker, A, Ahmed, G, Oldreive, C, Davis, Z, Richards, S, et al., Biallelic ATM Inactivation Significantly Reduces Survival in Patients Treated on the United Kingdom Leukemia Research Fund Chronic Lymphocytic Leukemia 4 Trial. Journal of Clinical Oncology, 2012. 30(36): p. 4524-32.

147. *Martínez-Trillos, A, Quesada, V, Villamor, N, Puente, XS, López-Otín, C, and Campo, E, Recurrent gene mutations in CLL, in Advances in Chronic Lymphocytic Leukemia. 2013, Springer. p. 87-107.*

148. *Puente, XS, Pinyol, M, Quesada, V, Conde, L, Ordóñez, GR, Villamor, N, et al., Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature, 2011. 475(7354): p. 101-05.*

149. *Quesada, V, Conde, L, Villamor, N, Ordonez, GR, Jares, P, Bassaganyas, L, et al., Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nature Genetics, 2012. 44(1): p. 47-52.*

150. *Wang, L, Lawrence, MS, Wan, Y, Stojanov, P, Sougnez, C, Stevenson, K, et al., SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia. New England Journal of Medicine, 2011. 365(26): p. 2497-506.*

151. *Zweidler-Mckay, PA, He, Y, Xu, L, Rodriguez, CG, Karnell, FG, Carpenter, AC, et al., Notch signaling is a potent inducer of growth arrest and apoptosis in a wide range of B-cell malignancies. Blood, 2005. 106(12): p. 3898-906.*

152. *Rosati, E, Sabatini, R, Rampino, G, Tabilio, A, Di Ianni, M, Fettucciari, K, et al., Constitutively activated Notch signaling is involved in survival and apoptosis resistance of B-CLL cells. Blood, 2009. 113(4): p. 856-65.*

153. *Villamor, N, Conde, L, Martinez-Trillos, A, Cazorla, M, Navarro, A, Bea, S, et al., NOTCH1 mutations identify a genetic subgroup of chronic lymphocytic leukemia patients with high risk of transformation and poor outcome. Leukemia, 2013. 27(5): p. 1100-06.*

154. *Mansouri, L, Cahill, N, Gunnarsson, R, Smedby, KE, Tjonnfjord, E, Hjalgrim, H, et al., NOTCH1 and SF3B1 mutations can be added to the hierarchical prognostic classification in chronic lymphocytic leukemia. Leukemia, 2013. 27(2): p. 512-14.*

155. *Fabbri, G, Rasi, S, Rossi, D, Trifonov, V, Khiabanian, H, Ma, J, et al., Analysis of the chronic lymphocytic leukemia coding genome: role of NOTCH1 mutational activation. The Journal of Experimental Medicine, 2011. 208(7): p. 1389-401.*

156. *David, CJ and Manley, JL, Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes & Development, 2010. 24(21): p. 2343-64.*

157. *G D Te Raa, IAMD, V Navrkalova, A Skowronska, P D Moerland, J van Laar, C Oldreive, H Monsuur, M Trbusek, J Malcikova, M Lodén, C H Geisler, J Hüllein, A Jethwa, T Zenz, S Pospisilova, T Stankovic, M H J van Oers, A P Kater and E Eldering, The impact of SF3B1 mutations in CLL on the DNA-damage response. Leukemia, 2014. 10(1038): p. 318.*

158. *Oscier, DG, Rose-Zerilli, MJJ, Winkelmann, N, Gonzalez De Castro, D, Gomez, B, Forster, J, et al., The clinical significance of NOTCH1 and SF3B1 mutations in the UK LRF CLL4 trial. Blood, 2013. 121(3): p. 468-75.*

159. *Rossi, D, Bruscaggin, A, Spina, V, Rasi, S, Khiabanian, H, Messina, M, et al., Mutations of the SF3B1 splicing factor in chronic lymphocytic leukemia: association with progression and fludarabine-refractoriness. Blood, 2011. 118(26): p. 6904-08.*

160. *Herishanu, Y, Pérez-Galán, P, Liu, D, Biancotto, A, Pittaluga, S, Vire, B, et al., The lymph node microenvironment promotes B-cell receptor signaling, NF-κB activation, and tumor proliferation in chronic lymphocytic leukemia. Blood, 2011. 117(2): p. 563-74.*

161. *Endo, T, Nishio, M, Enzler, T, Cottam, HB, Fukuda, T, James, DF, et al., BAFF and APRIL support chronic lymphocytic leukemia B-cell survival through activation of the canonical NF-κB pathway. Blood, 2007. 109(2): p. 703-10.*

162. *Rossi, D, Fangazio, M, Rasi, S, Vaisitti, T, Monti, S, Cresta, S, et al., Disruption of BIRC3 associates with fludarabine chemorefractoriness in TP53 wild-type chronic lymphocytic leukemia. Blood, 2012. 119(12): p. 2854-62.*

163. Rossi, D, Deaglio, S, Dominguez-Sola, D, Rasi, S, Vaisitti, T, Agostinelli, C, et al., Alteration of BIRC3 and multiple other NF-κB pathway genes in splenic marginal zone lymphoma. Blood, 2011. 118(18): p. 4930-34.

164. Compagno, M, Lim, WK, Grunn, A, Nandula, SV, Brahmachary, M, Shen, Q, et al., Mutations of multiple genes cause deregulation of NF-kB in diffuse large B-cell lymphoma. Nature, 2009. 459(7247): p. 717-21.

165. Puente, XS, Pinyol, M, Quesada, V, Conde, L, Ordóñez, GR, Villamor, N, et al., Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature, 2011. 475(7354): p. 101-05.

166. O'neill, LAJ and Bowie, AG, The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. Nature Review Immunology, 2007. 7(5): p. 353-64.

167. Burger, JA, Quiroga, MP, Hartmann, E, Bürkle, A, Wierda, WG, Keating, MJ, et al., High-level expression of the T-cell chemokines CCL3 and CCL4 by chronic lymphocytic leukemia B cells in nurselike cell cocultures and after BCR stimulation. Blood, 2009. 113(13): p. 3050-58.

168. Martínez-Trillos, A, Pinyol, M, Navarro, A, Aymerich, M, Jares, P, Juan, M, et al., Mutations in the Toll-like receptor/MYD88 pathway in chronic lymphocytic leukemia identify a subset of young patients with favorable outcome. Blood, 2014. 123(24): p.3790-96

169. Baliakas, P, Hadzidimitriou, A, Agathangelidis, A, Rossi, D, Sutton, L-A, Kminkova, J, et al., Prognostic relevance of MYD88 mutations in CLL: the jury is still out. Blood, 2015. 126(8): p. 1043-44.

170. Vaziri, H and Benchimol, S, Reconstitution of telomerase activity in normal human cells leads to elongation of telomeres and extended replicative life span. Current Biology, 1998. 8(5): p. 279-82.

171. Baumann, P and Price, C, Pot1 and telomere maintenance. FEBS Letters, 2010. 584(17): p. 3779-84.

172. Lin, TT, Letsolo, BT, Jones, RE, Rowson, J, Pratt, G, Hewamana, S, et al., Telomere dysfunction and fusion during the progression of chronic lymphocytic leukemia: evidence for a telomere crisis. Blood, 2010. 116(11): p. 1899-907.

173. Ramsay, AJ, Quesada, V, Foronda, M, Conde, L, Martinez-Trillos, A, Villamor, N, et al., POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. Nature Genetics, 2013. 45(5): p. 526-30.

174. Jacobs, JJL, Loss of Telomere Protection: Consequences and Opportunities. Frontiers in Oncology, 2013. 3: p. 88.

175. Clifford, R, Louis, T, Robbe, P, Ackroyd, S, Burns, A, Timbs, AT, et al., SAMHD1 is mutated recurrently in chronic lymphocytic leukemia and is involved in response to DNA damage. Blood, 2014. 123(7): p. 1021-31.

176. Rossi, D, SAMHD1: a new gene for CLL. Blood, 2014. 123(7): p. 951-52.

177. Rodríguez, D, Bretones, G, Quesada, V, Villamor, N, Arango, JR, López-Guillermo, A, et al., Mutations in CHD2 cause defective association with active chromatin in chronic lymphocytic leukemia. Blood, 2015. 126(2): p. 195-202.

178. Marfella, CGA, Ohkawa, Y, Coles, AH, Garlick, DS, Jones, SN, and Imbalzano, AN, Mutation of the SNF2 family member Chd2 affects mouse development and survival. Journal of Cellular Physiology, 2006. 209(1): p. 162-71.

179. Jones, JA and Byrd, JC, How will B-cell-receptor–targeted therapies change future CLL therapy? Blood, 2014. 123(10): p. 1455-60.

180. Slupsky, JR, Does B Cell Receptor Signaling in Chronic Lymphocytic Leukaemia Cells Differ from That in Other B Cell Types? Scientifica, 2014. 2014: ID: 208928. p. 14.

181. Yoshimura, M, Ishizawa, J, Ruvolo, V, Dilip, A, Quintás-Cardama, A, Mcdonnell, TJ, et al., Induction of p53-mediated transcription and apoptosis by exportin-1 (XPO1) inhibition in mantle cell lymphoma. Cancer Science, 2014. 105(7): p. 795-801.

182. Welcker, M and Clurman, BE, FBW7 ubiquitin ligase: a tumour suppressor at the crossroads of cell division, growth and differentiation. Nature Review Cancer, 2008. 8(2): p. 83-93.

183. Nakayama, KI and Nakayama, K, Ubiquitin ligases: cell-cycle control and cancer. Nature Review Cancer, 2006. 6(5): p. 369-81.

184. Akhoondi, S, Sun, D, Von Der Lehr, N, Apostolidou, S, Klotz, K, Maljukova, A, et al., FBXW7/hCDC4 Is a General Tumor Suppressor in Human Cancer. Cancer Research, 2007. 67(19): p. 9006-12.

185. Gianfelici, V, Activation of the NOTCH1 pathway in chronic lymphocytic leukemia. haematologica, 2012. 97(3): p. 328-30.

186. Giulia Fabbri, HK, Antony B. Holmes, Jiguang Wang,, Monica Messina, CGM, Laura Pasqualucci,, and Raul Rabadan, aRD-F, Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. Journal of Experimental Medicine, 2013. 210(11): p. 2273-88.

187. Fujimoto, K, Shibasaki, T, Yokoi, N, Kashima, Y, Matsumoto, M, Sasaki, T, et al., Piccolo, a Ca2+ Sensor in Pancreatic β-Cells: INVOLVEMENT OF cAMP-GEFII·Rim2·PICCOLO COMPLEX IN cAMP-DEPENDENT EXOCYTOSIS. Journal of Biological Chemistry, 2002. 277(52): p. 50497-502.

188. Hammond, CM, White, D, Tomic, J, Shi, Y, and Spaner, DE, Extracellular calcium sensing promotes human B-cell activation and function. Blood, 2007. 110(12): p. 3985-95.

189. Hammond, CM, Shi, Y, White, D, Cervi, D, Tomic, J, and Spaner, DE, The B-cell calcium sensor predicts progression of chronic lymphocytic leukemia. Leukemia, 2008. 23(2): p. 426-29.

190. Burns, A, Dreau, H, Hatton, C, Henderson, S, Taylor, J and Schuh, A. Targeted Gene Profiling Identifies Differential Genetic Make-up Depending On Chronic Lymphocytic Leukaemia Subtype. in 54 ASH Annual Meating and Exposition. 2012. Atlanta, GA: Ammerican Society of Haematology.

191. Fabbri, G, Khiabanian, H, Holmes, AB, Wang, J, Messina, M, Mullighan, CG, et al., Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. The Journal of Experimental Medicine, 2013. 210(11): p. 2273-88.

192. Lohr, JG, Stojanov, P, Lawrence, MS, Auclair, D, Chapuy, B, Sougnez, C, et al., Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proceedings of the National Academy of Sciences, 2012. 109(10): p. 3879-84.

193. Liu, C-X, Li, Y, Obermoeller-Mccormick, LM, Schwartz, AL, and Bu, G, The Putative Tumor Suppressor LRP1B, a Novel Member of the Low Density Lipoprotein (LDL) Receptor Family, Exhibits Both Overlapping and Distinct Properties with the LDL Receptor-related Protein. Journal of Biological Chemistry, 2001. 276(31): p. 28889-96.

194. Kan, Z, Zheng, H, Liu, X, Li, S, Barber, TD, Gong, Z, et al., Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. Genome Research, 2013. 23(9): p. 1422-33.

195. Ding, L, Getz, G, Wheeler, DA, Mardis, ER, Mclellan, MD, Cibulskis, K, et al., Somatic mutations affect key pathways in lung adenocarcinoma. Nature, 2008. 455(7216): p. 1069-75.

196. Cowin, PA, George, J, Fereday, S, Loehrer, E, Van Loo, P, Cullinane, C, et al., LRP1B Deletion in High-Grade Serous Ovarian Cancers Is Associated with Acquired

Chemotherapy Resistance to Liposomal Doxorubicin. Cancer Research, 2012. 72(16): p. 4060-73.

197. Alami, R, Fan, Y, Pack, S, Sonbuchner, TM, Besse, A, Lin, Q, et al., Mammalian linker-histone subtypes differentially affect gene expression in vivo. Proceedings of the National Academy of Sciences of the United States of America, 2003. 100(10): p. 5920-25.

198. Landau, DA, Carter, SL, Stojanov, P, Mckenna, A, Stevenson, K, Lawrence, MS, et al., Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell, 2013. 152(4): p. 714-26.

199. Manuylov, NL, Smagulova, FO, and Tevosian, SG, Fog2 excision in mice leads to premature mammary gland involution and reduced Esr1 gene expression. Oncogene, 2007. 26(36): p. 5204-13.

200. Stilgenbauer, S, Sander, S, Bullinger, L, Benner, A, Leupolt, E, Winkler, D, et al., Clonal evolution in chronic lymphocytic leukemia: acquisition of high-risk genomic aberrations associated with unmutated VH, resistance to therapy, and short survival. Haematologica, 2007. 92(9): p. 1242-45.

201. Ojha, J, Ayres, J, Secreto, C, Tschumper, R, Rabe, K, Van Dyke, D, et al., Deep sequencing identifies genetic heterogeneity and recurrent convergent evolution in chronic lymphocytic leukemia. Blood, 2015. 125(3): p. 492-98.

202. Lynda J Campbell, Cancer Cytogenetics: Methods and Protocols. Methods in Molecular Biology, ed. L.J. Campbell. 2011: Humanna Press. p. 273.

203. Wan, TSK and Ma, ESK, The role of FISH in hematologic cancer. International Journal of Hematologic Oncology, 2012. 1(1): p. 71-86.

204. Wan, TSK, Cancer Cytogenetics: Methodology Revisited. Annals of Laboratory Medicine, 2014. 34(6): p. 413-25.

205. Steemers, FJ, Chang, W, Lee, G, Barker, DL, Shen, R, and Gunderson, KL, Whole-genome genotyping with the single-base extension assay. Nature Methods, 2006. 3(1): p. 31-33.

206. Hagenkord, JM, Monzon, FA, Kash, SF, Lilleberg, S, Xie, Q, and Kant, JA, Array-Based Karyotyping for Prognostic Assessment in Chronic Lymphocytic Leukemia: Performance Comparison of Affymetrix 10K2.0, 250K Nsp, and SNP6.0 Arrays. The Journal of Molecular Diagnostics : Journal of Molecular Diagnostics, 2010. 12(2): p. 184-96.

207. Lehmann, S, Ogawa, S, Raynaud, SD, Sanada, M, Nannya, Y, Ticchioni, M, et al., Molecular allelokaryotyping of early-stage, untreated chronic lymphocytic leukemia. Cancer, 2008. 112(6): p. 1296-305.

208. Sathirapongsasuti, JF, Lee, H, Horst, BAJ, Brunner, G, Cochran, AJ, Binder, S, et al., Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics, 2011. 27(19): p. 2648-54.

209. Konstantinos, K, Panagiotis, P, Antonios, V, Agelos, P, and Argiris, N, PCR–SSCP: A Method for the Molecular Analysis of Genetic Diseases. Molecular Biotechnology, 2008. 38(2): p. 155-63.

210. Mahdieh, N and Rabbani, B, An Overview of Mutation Detection Methods in Genetic Disorders. Iranian Journal of Pediatrics, 2013. 23(4): p. 375-88.

211. Bräutigam, S, Kujat, A, Kirst, P, Seidel, J, Ümit Lüleyap, H, and Froster, UG, DHPLC mutation analysis of phenylketonuria. Molecular Genetics and Metabolism, 2003. 78(3): p. 205-10.

212. Wu, X-m, Fu, J-g, Ge, W-z, Zhu, J-y, Wang, J-y, Zhang, W, et al., Screen p53 mutations in hepatocellular carcinoma by FASAY: A novel splicing mutation. Journal of Zhejiang University. Science, 2007. 8(2): p. 81-87.

213. *Lin, K, Adamson, J, Johnson, GG, Carter, A, Oates, M, Wade, R, et al., Functional Analysis of the ATM-p53-p21 Pathway in the LRF CLL4 Trial: Blockade at the Level of p21 Is Associated with Short Response Duration. Clinical Cancer Research, 2012. 18(15): p. 4191-200.*

214. *Šmardová, J, Šmarda, J, and Koptíková, J, Functional analysis of p53 tumor suppressor in yeast. Differentiation, 2005. 73(6): p. 261-77.*

215. *Bakkar, AA, Quach, V, Le Borgne, A, Toublanc, M, Henin, D, Wallerand, H, et al., Sensitive Allele-Specific PCR Assay Able to Detect FGFR3 Mutations in Tumors and Urine from Patients with Urothelial Cell Carcinoma of the Bladder. Clinical Chemistry, 2005. 51(8): p. 1555-57.*

216. *Newton, CR, Graham, A, Heptinstall, LE, Powell, SJ, Summers, C, Kalsheker, N, et al., Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Research, 1989. 17(7): p. 2503-16.*

217. *Metzker, ML, Sequencing technologies [mdash] the next generation. Nature Review Genetics, 2010. 11(1): p. 31-46.*

218. *Metzker, ML, Emerging technologies in DNA sequencing. Genome Research, 2005. 15(12): p. 1767-76.*

219. *Liu, L, Li, Y, Li, S, Hu, N, He, Y, Pong, R, et al., Comparison of Next-Generation Sequencing Systems. Journal of Biomedicine and Biotechnology, 2012. 2012: p. 251364.*

220. *Flaherty, P, Natsoulis, G, Muralidharan, O, Winters, M, Buenrostro, J, Bell, J, et al., Ultrasensitive detection of rare mutations using next-generation targeted resequencing. Nucleic Acids Research, 2012. 40(1): p. e2-e2.*

221. *Van Dijk, EL, Auger, H, Jaszczyszyn, Y, and Thermes, C, Ten years of next-generation sequencing technology. Trends in Genetics, 2014. 30(9): p. 418-26.*

222. Grada, A and Weinbrecht, K, Next-Generation Sequencing: Methodology and Application. Journal of Investigative Dermatology, 2013. 133(8): p. e11.

223. Liu, L, Li, Y, Li, S, Hu, N, He, Y, Pong, R, et al., Comparison of Next-Generation Sequencing Systems. Journal of Biomedicine and Biotechnology, 2012. 2012: p. 11.

224. Michael a Quail, MS, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow and Yong Gu, A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics, 2012. 13(341). Doi:10.1186/1471-2164-13-341

225. Illuma. An introduction to next generation sequencing technology. 2016 Accessed Jan,2016.

226. Sheridan, C, Illumina claims [dollar]1,000 genome win. Nature Biotechnology, 2014. 32(2): p. 115-15.

227. Saunders, CJ, Miller, NA, Soden, SE, Dinwiddie, DL, Noll, A, Alnadi, NA, et al., Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units. Science Translational Medicine, 2012. 4(154): p. 154ra35-54ra35.

228. Ng, SB, Turner, EH, Robertson, PD, Flygare, SD, Bigham, AW, Lee, C, et al., Targeted Capture and Massively Parallel Sequencing of Twelve Human Exomes. Nature, 2009. 461(7261): p. 272-76.

229. Ljungström, V, Cortese, D, Young, E, Pandzic, T, Mansouri, L, Plevova, K, et al., Whole-exome sequencing in relapsing chronic lymphocytic leukemia: clinical impact of recurrent RPS15 mutations. Blood, 2016. 127(8): p. 1007-16.

230. Rehm, HL, Disease-targeted sequencing: a cornerstone in the clinic. Nature Review Genetics, 2013. 14(4): p. 295-300.

231. Sutton, L-A, Ljungström, V, Mansouri, L, Young, E, Cortese, D, Navrkalova, V, et al., Targeted next-generation sequencing in chronic lymphocytic leukemia: a high-throughput yet tailored approach will facilitate implementation in a clinical setting. Haematologica, 2015. 100(3): p. 370-76.

232. Mamanova, L, Coffey, AJ, Scott, CE, Kozarewa, I, Turner, EH, Kumar, A, et al., Target-enrichment strategies for next-generation sequencing. Nature Methods, 2010. 7(2): p. 111-18.

233. Perrott, J, Optimization and Improvement of Emulsion PCR for the Ion Torrent Next-Generation Sequencing Platform. 2011. p.15.

234. Mertes, F, Elsharawy, A, Sauer, S, Van Helvoort, JMLM, Van Der Zaag, PJ, Franke, A, et al., Targeted enrichment of genomic DNA regions for next-generation sequencing. Briefings in Functional Genomics, 2011. 10(6): p. 374-86.

235. Chang, F and Li, MM, Clinical application of amplicon-based next-generation sequencing in cancer. Cancer Genetics. 206(12): p. 413-19.

236. Crockett, DK, Genomic Applications in Pathology, 2015, New York: Springer ed. I. S. George Jabboure Netto. Doi 978-1-4939-0726-7.

237. Thorvaldsdóttir, H, Robinson, JT, and Mesirov, JP, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics, 2012. 14(12): p.178-92.

238. Bronner, IF, Quail, MA, Turner, DJ, and Swerdlow, H, Improved Protocols for Illumina Sequencing. Current protocols in Human Genetics / editorial board, Jonathan L. Haines ... [et al.], 2009. Doi: 10.1002/0471142905.hg1802s62.

239. Rothberg, JM, Hinz, W, Rearick, TM, Schultz, J, Mileski, W, Davey, M, et al., An integrated semiconductor device enabling non-optical genome sequencing. Nature, 2011. 475(7356): p. 348-52.

240. Wang, Y, Wen, Z, Shen, J, Cheng, W, Li, J, Qin, X, et al., Comparison of the performance of Ion Torrent chips in noninvasive prenatal trisomy detection. Journal of Human Genetics, 2014. 59(7): p. 393-96.

241. Baliakas, P, Hadzidimitriou, A, Sutton, LA, Rossi, D, Minga, E, Villamor, N, et al., Recurrent mutations refine prognosis in chronic lymphocytic leukemia. Leukemia, 2015. 29(2): p. 329-36.

242. Landau, DA, Carter, SL, Getz, G, and Wu, CJ, Clonal evolution in hematological malignancies and therapeutic implications. Leukemia, 2014. 28(1): p. 34-43.

243. Lin, K, Glenn, MA, Harris, RJ, Duckworth, AD, Dennett, S, Cawley, JC, et al., c-Abl Expression in Chronic Lymphocytic Leukemia Cells: Clinical and Therapeutic Implications. Cancer Research, 2006. 66(15): p. 7801-09.

244. Zenz, T, Habe, S, Denzel, T, Mohr, J, Winkler, D, Buhler, A, et al., Detailed analysis of p53 pathway defects in fludarabine-refractory chronic lymphocytic leukemia (CLL): dissecting the contribution of 17p deletion, TP53 mutation, p53-p21 dysfunction, and miR34a in a prospective clinical trial. Blood, 2009. 114(13): p. 2589-97.

245. Puente, XS, Pinyol, M, Quesada, V, Conde, L, Ordonez, GR, Villamor, N, et al., Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature, 2011. 475(7354): p. 101-05.

246. Burns, A, Dreau, H, Hatton, C, Henderson, S, Taylor, J and Schuh, A, Targeted Gene Profiling Identifies Differential Genetic Make-up Depending On Chronic Lymphocytic Leukaemia Subtype, in 54 Annual ASH Meeting and Exposition. 2012, American Society of Haematology: Georgia.

247. Villanueva-Cañas, JL, Laurie, S, and Albà, MM, Improving Genome-Wide Scans of Positive Selection by Using Protein Isoforms of Similar Length. Genome Biology and Evolution, 2013. 5(2): p. 457-67.

248. Baralle, D and Baralle, M, Splicing in action: assessing disease causing sequence changes. Journal of Medical Genetics, 2005. 42(10): p. 737-48.

249. Tewhey, R, Nakano, M, Wang, X, Pabón-Peña, C, Novak, B, Giuffre, A, et al., Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biology, 2009. 10(10): p. R116.

250. Sachidanandam, R, Weissman, D, Steven C. Schmidt, J M. Kakol & Lincoln D. Stein, A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature, 2001. 409(6822): p. 928-33.

251. Endo, TA, Quality control method for RNA-seq using single nucleotide polymorphism allele frequency. Genes to Cells, 2014. 19(11): p. 821-29.

252. Guo, Y, Long, J, He, J, Li, C-I, Cai, Q, Shu, X-O, et al., Exome sequencing generates high quality data in non-target regions. BMC Genomics, 2012. 13(1): p. 1-10.

253. Durbin Rm, AD, Abecasis GR, et al., A map of human genome variation from population-scale sequencing. Nature, 2010. 467(7319): p. 1061-73.

254. Guo, Y, Ye, F, Sheng, Q, Clark, T, and Samuels, DC, Three-stage quality control strategies for DNA re-sequencing data. Briefings in Bioinformatics, 2014. 15(6): p. 879-89.

255. Armbruster, DA and Pry, T, Limit of blank, limit of detection and limit of quantitation. Clinical Biochemist Review, 2008. 29(Suppl 1): p. S49-52.

256. Pisareva, E, Gutkina, N, Kovalenko, S, and Shamanin, V, P65: Development of allele-specific PCR assays for detection of mutations in KRAS gene in colorectal cancer in Russian patients. European Journal of Cancer Supplements, 2015. 13(1): p. 42-43.

257. Rehm, HL, Bale, SJ, Bayrak-Toydemir, P, Berg, JS, Brown, KK, Deignan, JL, et al., ACMG clinical laboratory standards for next-generation sequencing. Genetic Medicine, 2013. 15(9): p. 733-47.

258. Frampton, GM, Fichtenholtz, A, Otto, GA, Wang, K, Downing, SR, He, J, et al., Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nature Biotechnology, 2013. 31(11): p. 1023-31.

259. Hurtado, AM, Chen-Liang, TH, Przychodzen, B, Hamedi, C, Muñoz-Ballester, J, Dienes, B, et al., Prognostic signature and clonality pattern of recurrently mutated genes in inactive chronic lymphocytic leukemia. Blood Cancer Journal, 2015. 5(8): p. e342.

260. Sundaresan, TK and Haber, DA, Does molecular monitoring matter in early-stage breast cancer? Science Translational Medicine, 2015. 7(302): p. 302fs35-02fs35.

261. Rabbani, B, Tekin, M, and Mahdieh, N, The promise of whole-exome sequencing in medical genetics. Journal of Humuman Genetics, 2014. 59(1): p. 5-15.

262. Chen, Y-C, Liu, T, Yu, C-H, Chiang, T-Y, and Hwang, C-C, Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. PLoS ONE, 2013. 8(4): p. e62856.

263. Valencia, CA, Rhodenizer, D, Bhide, S, Chin, E, Littlejohn, MR, Keong, LM, et al., Assessment of Target Enrichment Platforms Using Massively Parallel Sequencing for the Mutation Detection for Congenital Muscular Dystrophy. Journal of Molecular Diagnostics, 2012. 14(3): p. 233-46.

264. Bodi, K, Perera, AG, Adams, PS, Bintzler, D, Dewar, K, Grove, DS, et al., Comparison of Commercially Available Target Enrichment Methods for Next-Generation Sequencing. Journal of Biomolecular Techniques, 2013. 24(2): p. 73-86.

265. Reumers, J, De Rijk, P, Zhao, H, Liekens, A, Smeets, D, Cleary, J, et al., Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. Nature Biotechnology, 2012. 30(1): p. 61-68.

266. Huang, D, Kim, D-W, Kotsakis, A, Deng, S, Lira, P, Ho, SN, et al., Multiplexed deep sequencing analysis of ALK kinase domain identifies resistance mutations in relapsed patients following crizotinib treatment. Genomics, 2013. 102(3): p. 157-62.

267. Pleasance, ED, Stephens, PJ, O'meara, S, Mcbride, DJ, Meynert, A, Jones, D, et al., A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature, 2010. 463(7278): p. 184-90.

268. Knouse, KA, Wu, J, and Amon, A, Assessment of megabase-scale somatic copy number variation using single-cell sequencing. Genome Research, 2016. 26(3): p. 376-84.

269. Macdonald, JR, Ziman, R, Yuen, RKC, Feuk, L, and Scherer, SW, The database of genomic variants: a curated collection of structural variation in the human genome. Nucleic Acids Research, 2013. 42: p. D986–D92.

270. Mckusick, VA, Mendelian Inheritance in Man and Its Online Version, OMIM. The American Journal of Human Genetics, 2007. 80(4): p. 588-604.

271. Huret, J-L, Ahmad, M, Arsaban, M, Bernheim, A, Cigna, J, Desangles, F, et al., Atlas of Genetics and Cytogenetics in Oncology and Haematology in 2013. Nucleic Acids Research, 2013. 41(Database issue): p. D920-D24.

272. Mansouri, L, Sutton, L-A, Ljungström, V, Bondza, S, Arngården, L, Bhoi, S, et al., Functional loss of IκBε leads to NF-κB deregulation in aggressive chronic lymphocytic leukemia. The Journal of Experimental Medicine, 2015. 212(6): p. 833-43.

273. Nadeu, F, Delgado, J, Royo, C, Baumann, T, Stankovic, T, Pinyol, M, et al., Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1 and ATM mutations in chronic lymphocytic leukemia. Blood, 2016. 127(17): p. 2122-30.

274. Jean Fan, LW, Angela N Brooks, Youzhong Wan, Donna S Neuberg, Laura Z. Rassenti, Emanuela M. Ghia, Thomas J. Kipps, et al, Comprehensive Bulk and Single Cell Transcriptomic Characterization of SF3B1 Mutation Reveals Its Pleiotropic Effects in Chronic Lymphocytic Leukemia, in ASH 57th Annual Meeting and Exposition. 2015: Orlando, Florida (126). p. 23.

275. Vollbrecht, C, Mairinger, FD, Koitzsch, U, Peifer, M, Koenig, K, Heukamp, LC, et al., Comprehensive Analysis of Disease-Related Genes in Chronic Lymphocytic Leukemia by Multiplex PCR-Based Next Generation Sequencing. PLoS ONE, 2015. 10(6): p. e0129544.

276. Korbel, Jan O and Campbell, Peter J, Criteria for Inference of Chromothripsis in Cancer Genomes. Cell, 2013. 152(6): p. 1226-36.

277. Schuh A, BJ, Humphray S, Alexa A, Burns A, Clifford R, Stephan M. Feller, et al, Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. Blood, 2012. 120(20): p. 4191-96.

278. Puente, XS, Bea, S, Valdes-Mas, R, Villamor, N, Gutierrez-Abril, J, Martin-Subero, JI, et al., Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature, 2015. 526(7574): p. 519-24.

279. Messina, M, Del Giudice, I, Khiabanian, H, Rossi, D, Chiaretti, S, Rasi, S, et al., Genetic lesions associated with chronic lymphocytic leukemia chemo-refractoriness. Blood, 2014. 123(15): p. 2378-88.

280. Ljungström, V, Cortese, D, Young, E, Pandzic, T, Mansouri, L, Plevova, K, et al., Whole-exome sequencing in relapsing chronic lymphocytic leukemia: clinical impact of recurrent RPS15 mutations. Blood, 2016. 127(8): p. 1007-16.

281. Wan, Y and Wu, CJ, SF3B1 mutations in chronic lymphocytic leukemia. Blood, 2013. 121(23): p. 4627-34.

282. Martínez-Trillos, A, Pinyol, M, Navarro, A, Aymerich, M, Jares, P, Juan, M, et al., Mutations in TLR/MYD88 pathway identify a subset of young chronic lymphocytic leukemia patients with favorable outcome. Blood, 2014. 123(24): p. 3790-96.

283. Dal Bo, M, Rossi, FM, Rossi, D, Deambrogi, C, Bertoni, F, Del Giudice, I, et al., 13q14 Deletion size and number of deleted cells both influence prognosis in chronic lymphocytic leukemia. Genes, Chromosomes and Cancer, 2011. 50(8): p. 633-43.

284. Orlandi, EM, Bernasconi, P, and Pascutto, C, The prognostic difference of monoallelic versus biallelic deletion of 13q in chronic lymphocytic leukemia. Cancer, 2012. 118(20): p. 5179-79.

285. Ramsay, AJ, Rodriguez, D, Villamor, N, Kwarciak, A, Tejedor, JR, Valcarcel, J, et al., Frequent somatic mutations in components of the RNA processing machinery in chronic lymphocytic leukemia. Leukemia, 2013. 27(7): p. 1600-03.

286. Ouillette, P, Saiya-Cork, K, Seymour, E, Li, C, Shedden, K, and Malek, SN, Clonal evolution, genomic drivers and effects of therapy in chronic lymphocytic leukemia. Clinical cancer research : an official journal of the American Association for Cancer Research, 2013. 19(11): p. 2893-904.

287. Alexandrov, LB and Stratton, MR, Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Current Opinion in Genetics & Development, 2014. 24(100): p. 52-60.

288. Alexandrov, LB, Nik-Zainal, S, Wedge, DC, Aparicio, SAJR, Behjati, S, Biankin, AV, et al., Signatures of mutational processes in human cancer. Nature, 2013. 500(7463): p. 415-21.

289. Dollé, MET, Snyder, WK, Dunson, DB, and Vijg, J, Mutational fingerprints of aging. Nucleic Acids Research, 2002. 30(2): p. 545-49.

290. *Kasar, S, Kim, J, Improgo, R, Tiao, G, Polak, P, Haradhvala, N, et al., Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. Nature Communications, 2015. 7(6): p. 8866-76.*

291. *Foà, R, Del Giudice, I, Guarini, A, Rossi, D, and Gaidano, G, Clinical implications of the molecular genetics of chronic lymphocytic leukemia. Haematologica, 2013. 98(5): p. 675-85.*

292. *Petitjean, A, Achatz, MIW, Borresen-Dale, AL, Hainaut, P, and Olivier, M, TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. Oncogene, 2007. 26(15): p. 2157-65.*

293. *Rossi, D, Khiabanian, H, Rasi, S, Ciardullo, C, Terzi Di Bergamo, L, Monti, S, et al., Small Subclones Harboring NOTCH1, SF3B1 or BIRC3 Mutations Are Clinically Irrelevant in Chronic Lymphocytic Leukemia. Blood, 2014. 124(21): p. 295-95.*

294. *Keim, C, Kazadi, D, Rothschild, G, and Basu, U, Regulation of AID, the B-cell genome mutator. Genes & Development, 2013. 27(1): p. 1-17.*

295. *Fox, EJ, Salk, JJ, and Loeb, LA, Exploring the implications of distinct mutational signatures and mutation rates in aging and cancer. Genome Medicine, 2016. 8(1): p. 1-3.*

296. *Plander, M, Seegers, S, Ugocsai, P, Diermeier-Daucher, S, Ivanyi, J, Schmitz, G, et al., Different proliferative and survival capacity of CLL-cells in a newly established in vitro model for pseudofollicles. Leukemia, 2009. 23(11): p. 2118-28.*

297. *Rebhandl, S, Huemer, M, Gassner, FJ, Zaborsky, N, Hebenstreit, D, Catakovic, K, et al., APOBEC3 signature mutations in chronic lymphocytic leukemia. Leukemia, 2014. 28(9): p. 1929-32.*

298. Te Raa, GD, Derks, IAM, Navrkalova, V, Skowronska, A, Moerland, PD, Van Laar, J, et al., The impact of SF3B1 mutations in CLL on the DNA-damage response. Leukemia, 2015. 29(5): p. 1133-42.

299. Schwaederlé, M, Ghia, E, Rassenti, LZ, Obara, M, Dell' Aquila, ML, Fecteau, JF, et al., Subclonal evolution involving SF3B1 mutations in chronic lymphocytic leukemia. Leukemia, 2013. 27(5): p. 1214-17.

300. Nabhan, C, Raca, G, and Wang, Y, PRedicting prognosis in chronic lymphocytic leukemia in the contemporary era. JAMA Oncology, 2015. 1(7): p. 965-74.

301. Beerenwinkel, N, Schwarz, RF, Gerstung, M, and Markowetz, F, Cancer Evolution: Mathematical Models and Computational Inference. Systematic Biology, 2015. 64(1): p. e1-e25.

302. Rose-Zerilli, MJJ, Gibson, J, Wang, J, Tapper, W, Davis, Z, Parker, H, et al., Longitudinal copy number, whole exome and targeted deep sequencing of 'good risk' IGHV-mutated CLL patients with progressive disease. Leukemia, 2016. 30(6): p. 1301-10.

303. Rozovski, U, Hazan-Halevy, I, Keating, MJ, and Estrov, Z, Personalized medicine in CLL: Current status and future perspectives. Cancer Letters, 2014. 352(1): p. 4-14.

304. Biesecker, LG, Shianna, KV, and Mullikin, JC, Exome sequencing: the expert view. Genome Biology, 2011. 12(9): p. 128-28.

305. Hedegaard, J, Thorsen, K, Lund, MK, Hein, A-MK, Hamilton-Dutoit, SJ, Vang, S, et al., Next-Generation Sequencing of RNA and DNA Isolated from Paired Fresh-Frozen and Formalin-Fixed Paraffin-Embedded Samples of Human Cancer and Normal Tissue. PLoS ONE, 2014. 9(5): p. e98187.

306. Ulahannan, D, Kovac, MB, Mulholland, PJ, Cazier, JB, and Tomlinson, I, Technical and implementation issues in using next-generation sequencing of cancers in clinical practice. British Journal of Cancer, 2013. 109(4): p. 827-35.

307. Klco, JM, Spencer, DH, Miller, CA, Griffith, M, Lamprecht, TL, O'laughlin, M, et al., Functional heterogeneity of genetically defined subclones in acute myeloid leukemia. Cancer Cell, 2014. 25(3): p. 379-92.

308. Dienstmann, R, Dong, F, Borger, D, Dias-Santagata, D, Ellisen, LW, Le, LP, et al., Standardized decision support in next generation sequencing reports of somatic cancer variants. Molecular Oncology, 2014. 8(5): p. 859-73.

309. Quail, MA, Smith, M, Coupland, P, Otto, TD, Harris, SR, Connor, TR, et al., A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics, 2012. 13(1): p. 1-13.

310. Damm, F, Mylonas, E, Cosson, A, Yoshida, K, Della Valle, V, Mouly, E, et al., Acquired Initiating Mutations in Early Hematopoietic Cells of CLL Patients. Cancer Discovery, 2014. 4(9): p. 1088-101.

311. Kantorova, B, Malcikova, J, Navrkalova, V, Smardova, J, Brazdilova, K, Plevova, K, et al., Prognostic Impact of NOTCH1 Hotspot Mutation in TP53-Mutated Patients with Chronic Lymphocytic Leukemia. Blood, 2014. 124(21): p. 3283-83.

312. Henderson, S and Fenton, T, APOBEC3 genes: retroviral restriction factors to cancer drivers. Trends in Molecular Medicine. 21(5): p. 274-84.

313. Stephens, PJ, Greenman, CD, Fu, B, Yang, F, Bignell, GR, Mudie, LJ, et al., Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. Cell, 2011. 144(1): p. 27-40.

314. Forment, JV, Kaidi, A, and Jackson, SP, Chromothripsis and cancer: causes and consequences of chromosome shattering. Nature Review Cancer, 2012. 12(10): p. 663-70.

315. Nadeu, F, Delgado, J, Royo, C, Baumann, T, Stankovic, T, Pinyol, M, et al., Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1, and ATM mutations in chronic lymphocytic leukemia. Blood, 2016. 127(17): p. 2122-30.

316. Rai, KR and Jain, P, Chronic lymphocytic leukemia (CLL)Then and now. American Journal of Hematology, 2016. 91(3): p. 330-41.

317. Sabina, C, Marilisa, M, Ilaria, DG, Silvia, B, Alfonso, P, Monica, M, et al., NOTCH1, SF3B1, BIRC3 and TP53 mutations in patients with chronic lymphocytic leukemia undergoing first-line treatment: correlation with biological parameters and response to treatment. Leukemia & lymphoma, 2014. 55(12): p. 2785-92.

318. Liu, T-M, Woyach, JA, Zhong, Y, Lozanski, A, Lozanski, G, Dong, S, et al., Hypermorphic mutation of phospholipase C, γ2 acquired in ibrutinib-resistant CLL confers BTK independency upon B-cell receptor activation. Blood, 2015. 126(1): p. 61-68.

319. Fridman, S. and Lowe, S, Control of apoptosis by p53. Oncogene, 2003. (22): P. 9030–9040.

320. Jin S & Levine AJ. The p53 functional circuit. J. Cell Sci. (2001). 114: P.4139−4140

321. *Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., & Quince, C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform Nucleic Acids Research,* (2015). *43(6), e37. http://doi.org/10.1093/nar/gku1341*

322. *Lawlley, D. Historical origins of current concepts of carcinogenesis.Adv Cancer Res. (1994). 65(17) p.111.*

323. *Jesse, J., Edward, J. and Lawrence A. Mutation heterogeniety in cancer: origin and consequences. Annu Rev. Pathol. (2010). Doi 10. 1146.*